215
AUG 4 1987
COMPLETED 330-E
74.
D 101.60/4:716

Technical Report 716

Project A:
Improving the Selection, Classification, and Utilization of Army Enlisted Personnel

Development and Field Test of Army-Wide Rating Scales and the Rater Orientation and Training Program

Elaine D. Pulakos and Walter C. Borman, Editors
Personnel Decisions Research Institute

Selection and Classification Technical Area
Manpower and Personnel Research Laboratory

U.S. Army

Research Institute for the Behavioral and Social Sciences

July 1986

ORIGINAL

Approved for public release; distribution unlimited.

BEST COPY AVAILABLE

U. S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

WM. DARRYL HENDERSON
Colonel, IN
Commander

Research accomplished under contract
to the Department of the Army

Human Resources Research Organization

Technical review by

Leonard White
Darlene Olson

Notices

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences. ATTN: PERI-TST, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

BEST COPY AVAILABLE

REPORT DOCUMENTATION PAGE

1a REPORT SECURITY CLASSIFICATION: UNCLASSIFIED	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution unlimited.
2b DECLASSIFICATION/DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S)	5 MONITORING ORGANIZATION REPORT NUMBER(S): ARI Technical Report 716
6a NAME OF PERFORMING ORGANIZATION: Human Resources Research Organization; 6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION: U.S. Army Research Institute for the Behavioral and Social Sciences
6c ADDRESS (City, State, and ZIP Code): 1100 South Washington Street, Alexandria, VA 22314-4499	7b ADDRESS (City, State, and ZIP Code): 5001 Eisenhower Avenue, Alexandria, VA 22333-5600
8a NAME OF FUNDING/SPONSORING ORGANIZATION: U.S. Army Research Institute for the Behavioral and Social Sciences; 8b OFFICE SYMBOL (If applicable): PERI-RS	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: MDA 903-82-C-0531
8c ADDRESS (City, State, and ZIP Code): 5001 Eisenhower Avenue, Alexandria, VA 22333-5600	10 SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO: 20263731; PROJECT NO: A792; TASK NO; WORK UNIT ACCESSION NO.

11 TITLE (Include Security Classification)
Development and Field Test of Army-Wide Rating Scales and the Rater Orientation and Training Program

12 PERSONAL AUTHOR(S)
Pulakos, Elaine D., & Borman, Walter C., Editors, Personnel Decisions Research Institute

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
	FROM Oct 83 TO Oct 85	July 1986	74

16 SUPPLEMENTARY NOTATION
Lawrence M. Hanser, Contracting Officer's Representative.

17 COSATI CODES			18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Army-Wide Measures, Classification, Criterion Measures, First-Term Evaluation, Predictor Measures, Project A Field Test, Rate Training, Rating Scales, Selection, Soldier Effectiveness

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

The research described in this report was performed under Project A, the U.S. Army's current, large-scale manpower and personnel effort for improving the selection, classification, and utilization of Army enlisted personnel. This research sought to develop dimensions of soldier performance for evaluating first-term soldiers in any Military Occupational Specialty (MOS). A preliminary conceptual model of soldier effectiveness helped guide development of empirical scales. Behavioral analysis was used to identify and define effectiveness dimensions. The Skill Level I Common Task Soldier's Manual guided development of another set of rating scales in several task areas involving all first-term soldiers. A rater training program was prepared to help peers and supervisors make accurate evaluations using the Army-wide scales.

The rating scales and training program were field tested for nine MOS. A total of 904 supervisor and 1,206 peer raters evaluated a total of 1,369 first-term soldiers. Results were
(continued)

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION	
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	UNCLASSIFIED	
22a NAME OF RESPONSIBLE INDIVIDUAL: Lawrence M. Hanser	22b TELEPHONE (Include Area Code): 202/274-8275	22c OFFICE SYMBOL: PERI-RS

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted; All other editions are obsolete

BEST COPY AVAILABLE

ARI Technical Report 716

19. (Continued)

encouraging. Raters appeared to understand and comply with instruction, rating distributions were acceptable, and interrater reliabilities were reasonably high. The field tests also provided information that guided refinement of both the rating scales and the rater training program.

The appendixes that present additional documentation for this research are contained in a separate report with limited distribution (ARI Research Note 87-22, April 1987): Development and Test of Army-Wide Rating Scales and the Rater Orientation and Training Program: Appendixes to ARI Technical Report 716.

BEST COPY AVAILABLE

Technical Report 716

Project A:
Improving the Selection, Classification, and Utilization of Army Enlisted Personnel

Development and Field Test of Army-Wide Rating Scales and the Rater Orientation and Training Program

Elaine D. Pulakos and Walter C. Borman, Editors
Personnel Decisions Research Institute

Selection and Classification Technical Area
Lawrence M. Hanser, Chief

Manpower and Personnel Research Laboratory
Newell K. Eaton, Director

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

July 1986

Army Project Number
2Q263731A792

Manpower and Personnel

Approved for public release; distribution unlimited

BEST COPY AVAILABLE

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

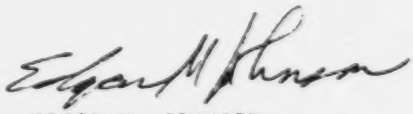
BEST COPY AVAILABLE

FOREWORD

The U.S. Army is currently making a large-scale effort in the manpower and personnel area to improve the selection, classification, and utilization of Army enlisted personnel. This document describes the development and field testing of Army-wide rating scales for evaluating performance of first-term enlisted personnel.

Impetus for the project came from the practical, professional, and legal need to validate ASVAB (the Armed Services Vocational Aptitude Battery--the current U.S. military selection/classification test battery) and other selection variables as predictors of training performance.

The portion of the effort described herein--"Project A"--is devoted to the development and validation of Army Selection and Classification Measures. Another part of the effort--"Project B"--is the development of a prototype Computerized Personnel Allocation System. Together, these Army Research Institute research efforts, with their in-house and contract components, compose a major program to develop a state-of-the-art, empirically validated system for personnel selection, classification, and allocation.



EDGAR M. JOHNSON
Technical Director

BEST COPY AVAILABLE

ACKNOWLEDGMENTS

Both editors contributed substantially to compiling this report; participation in performing the research is denoted by the individual chapter authorships. The authors thank Len White and Darlene Olson of the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) for their thoughtful comments on an earlier draft of this report. Steve Motowidlo of Personnel Decisions Research Institute (PDRI), Larry Hanser and Mike Rumsey of ARI, and Barry Reigelhaupt and Carolyn DeMeyer Harris of the Human Resources Research Organization (HumRRO) provided skilled assistance at various stages of the Army-wide scale development work. Barry Reigelhaupt and Jim Harris of HumRRO and Darlene Olson, Len White, and Larry Hanser of ARI provided invaluable help with the behavioral analysis workshops. A number of individuals participated in the field test rating scale administrations and provided helpful suggestions for improving the rater orientation and training program. These individuals included: Mike Bosshardt, VyVy Corpe, Steve Lammlein, Jeff McHenry, Teresa Russell, and Jody Toquam of PDRI; Jim Harris and Barry Reigelhaupt of HumRRO; Laurie Wise of the American Institutes for Research (AIR); and Len White, Mike Rumsey, and Darlene Olson of ARI. Laurie Wise and Winnie Young of AIR were responsible for compiling the data base and performing most of the analyses reported here, and Matt McGue of PDRI also assisted with the data analyses.

BEST COPY AVAILABLE

DEVELOPMENT AND FIELD TEST OF ARMY-WIDE RATING SCALES AND THE RATER ORIENTATION AND TRAINING PROGRAM

EXECUTIVE SUMMARY

Requirements:

Project A is a large-scale, multiyear research program intended to improve the selection and classification system for initial assignment of persons to U.S. Army Military Occupational Specialties (MOS). Experimental predictors (e.g., an interest inventory, computerized perceptual tests) are being developed to forecast job performance in the different MOS. To assess the validity of these predictors, special performance measures are also being developed.

This report describes the development and field tests of Army-wide rating scales, dimensions of performance that are relevant to first-term soldiers in any MOS. Also described in this report are the development and field tests of a rater orientation and training program intended to help peer and supervisor raters provide accurate evaluations using the Army-wide scales.

Procedure:

A preliminary model of soldier effectiveness helped guide development efforts. The behavior analysis method was employed to identify soldier effectiveness dimensions and to develop behavioral definitions of performance for each dimension. The resulting Army-wide behavioral rating scales were then readied for field testing with supervisor and peer raters in nine different MOS. In addition, the Skill Level I Common Task Soldier's Manual guided development of another set of rating scales intended to measure performance in several task areas for which all first-tour soldiers are responsible. These scales were also readied for field testing. Finally, a rater orientation and training program was prepared to help supervisor and peer raters make their performance ratings as accurate as possible. This program was also tried out in the field tests.

The research staff conducted two separate field tests. The first (Batch A) field test focused on four MOS, and the second (Batch B) field test focused on five other MOS. A total of 904 supervisors and 1,206 peer raters participated in the field tests, evaluating 1,369 first-term soldiers in all.

Results:

Results of the field tests were encouraging. In particular, (1) rater participants seemed accepting of the rater orientation and training program and appeared to understand and comply with the instructions; (2) the rating distributions were acceptable, with means a little above the scale mid-points and standard deviations comparable to those found in other validation research;

BEST COPY AVAILABLE

and (3) the interrater reliabilities were acceptably high, for both peer and supervisor raters.

In addition, an experiment was conducted during the Batch B field tests to assess the effects of rating practice on interrater reliability, rating errors, and rating accuracy. The results of the experiment showed practice to have no effect on any of the dependent measures.

Utilization of Findings:

The Army-wide rating scales will be used in Project A's Concurrent Validation to provide one set of criterion scores against which the validity of predictor measures can be assessed. The field tests described in this report provided valuable information that guided revision and further refinement of both the rating scales and the rater orientation and training program for use in the Concurrent Validation effort.

BEST COPY AVAILABLE

DEVELOPMENT AND FIELD TEST OF ARMY-WIDE RATING SCALES AND THE RATER ORIENTATION AND TRAINING PROGRAM

CONTENTS

	Page
OVERVIEW OF PROJECT A	1
CHAPTER 1: RATIONALE FOR THE ARMY-WIDE RATING SCALES	
Introduction	4
A Model of Soldier Effectiveness	5
CHAPTER 2: DEVELOPMENT OF THE ARMY-WIDE RATING SCALES AND TASK DIMENSIONS	
Summary of Procedures	8
Army-Wide Rating Scales Behavioral Analysis Workshops	8
Retranslation of the Behavioral Examples	13
Retranslation Results	13
Development of the Army-Wide Common Task Dimensions	15
CHAPTER 3: FIELD TEST PROCEDURES FOR ADMINISTERING THE ARMY-WIDE RATING SCALES	
Summary of Procedures	17
Sample and Procedures	17
CHAPTER 4: ARMY-WIDE RATING DATA ANALYSES AND RESULTS	
Data Analyses	22
Identification of Outlier Raters	24
Results	25
Revision of the Army-Wide Scales	31
CHAPTER 5: RATER ORIENTATION AND TRAINING	
Overview of the Approach	33
Lessons Learned During the Batch A Field Tests	35
Lessons Learned During the Batch B Field Tests	37
CHAPTER 6: BATCH B RATER TRAINING EXPERIMENT: THE EFFECTS OF PRACTICE ON MAKING RATINGS	
Purpose of the Experiment	43
Method	43
Training Programs	46
Development of Vignettes to Assess Accuracy	48
Dependent Variables	48
Results	51
Summary and Conclusions	54
GENERAL SUMMARY AND CONCLUSIONS	58
REFERENCES	59

BEST COPY AVAILABLE

CONTENTS (Continued)

Page

APPENDIXES*

APPENDIX A. ORIENTATION MATERIALS FOR BEHAVIORAL ANALYSIS WORKSHOPS A-1

B. RETRANSLATION MATERIALS AND EDITED BEHAVIORAL EXAMPLES B-1

C. FIELD TEST VERSION OF THE ARMY-WIDE BEHAVIOR-BASED AND COMMON TASK RATING SCALES C-1

D. ADMINISTRATOR'S INSTRUCTIONS: IDENTIFICATION OF POTENTIAL PEER RATERS D-1

E. MEANS, STANDARD DEVIATIONS, INTERRATER RELIABILITIES, AND INTERCORRELATIONS FOR THE ARMY-WIDE BEHAVIORAL DIMENSIONS BY MOS E-1

F. MEANS, STANDARD DEVIATIONS, INTERRATER RELIABILITIES, AND INTERCORRELATIONS FOR THE ARMY-WIDE COMMON TASK DIMENSIONS BY MOS F-1

G. CORRELATIONS BETWEEN SELECTED RATING MEASURES BY MOS G-1

H. CONCURRENT VALIDATION VERSION OF THE ARMY-WIDE BEHAVIOR-BASED AND COMMON TASK RATING SCALES H-1

I. ADMINISTRATOR'S MANUAL: PEER AND SUPERVISOR RATING SESSIONS I-1

J. VIGNETTE RATING TASK FOR PEERS AND SUPERVISORS J-1

LIST OF TABLES

Table 1. Participants in Behavior Analysis Workshops 9

2. Soldier Effectiveness Examples Generated 12

3. Number of Behavioral Examples Reliably Retranslated into Each Dimension 14

4. Number of Soldiers in the Field Tests by MOS and Location . . 18

*NOTE: The Appendixes (A-J) to this report are contained in ARI Research Note 87-22 (April 1987).

BEST COPY AVAILABLE

Page

LIST OF TABLES (Continued)

Table 5. Soldiers in the Field Tests by Sex and Race 19

6. Rater and Ratee Sample Sizes in the Field Tests 21

7. Frequency Distributions of Ratings Across the Seven Rating Scale Points (in Percents) 26

8. Means and Standard Deviations (SD) of Selected Army-Wide Measures 27

9. Intraclass Correlations Coefficients for Selected Army-Wide Measures 28

10. Correlations Between Rating Measures Averaged Across All MOS 30

11. Breakdown of the Rater Training Sample by Location and MOS 44

12. True Score Matrix for Vignette Ratees on Six Army-Wide Dimensions 49

13. Interrater Reliabilities by Training Condition Across All MOS 52

14. Rating Error Dependent Variables by Training Condition and Rater Group: Army-Wide Behavioral Rating Scales 55

15. Rating Error Dependent Variables by Training Condition and Rater Group: MOS-Specific Behavioral Rating Scales . . . 56

16. Accuracy Means by Training Condition and Rater Group 57

LIST OF FIGURES

Figure 1. A Preliminary Model of Soldier Effectiveness 7

2. Pictorial Depiction of Halo Error 38

3. Pictorial Depiction of Same-Level-of-Effectiveness Error . . 39

4. Critical Data Verification Checks: Rating Sessions 41

5. Supervisor Practice Vignette 47

BEST COPY AVAILABLE

OVERVIEW OF PROJECT A

Project A is a comprehensive long-range research and development program which the U.S. Army has undertaken to develop an improved personnel selection and classification system for enlisted personnel. The Army's goal is to increase its effectiveness in matching first-tour enlisted manpower requirements with available personnel resources, through use of new and improved selection/classification tests which will validly predict carefully developed measures of job performance. The project addresses the 675,000-person enlisted personnel system of the Army, encompassing several hundred different military occupations.

This research program began in 1980, when the U.S. Army Research Institute (ARI) started planning the extensive research effort that would be needed to develop the desired system. In 1982 a consortium led by the Human Resources Research Organization (HumRRO) and including the American Institutes for Research (AIR) and the Personnel Decisions Research Institute (PDRI) was selected by ARI to undertake the 9-year project. The total project utilizes the services of 40 to 50 ARI and consortium researchers working collegially in a variety of specialties, such as industrial and organizational psychology, operations research, management science, and computer science.

The specific objectives of Project A are to:

- Validate existing selection measures against both existing and project-developed criteria. The latter are to include both Army-wide job performance measures based on newly developed rating scales, and direct hands-on measures of MOS-specific task performance.
- Develop and validate new selection and classification measures.
- Validate intermediate criteria (e.g., performance in training) as predictors of later criteria (e.g., job performance ratings), so that better informed reassignment and promotion decisions can be made throughout a soldier's career.
- Determine the relative utility to the Army of different performance levels across MOS.
- Estimate the relative effectiveness of alternative selection and classification procedures in terms of their validity and utility for making operational selection and classification decisions.

The research design for the project incorporates three main stages of data collection and analysis in an iterative progression of development, testing, evaluation, and further development of selection/classification instruments (predictors) and measures of job performance (criteria). In the first iteration, file data from Army accessions in fiscal years (FY) 1981 and 1982 were evaluated to explore the relationships between the scores of applicants on the Armed Services Vocational Aptitude Battery (ASVAB), and their subsequent performance in training and their scores on the first-tour Skills and Qualification Tests (SQT).

BEST COPY AVAILABLE

In the second iteration, a concurrent validation design will be executed with FY83/84 accessions. As part of the preparation for the Concurrent Validation, a "preliminary battery" of perceptual, spatial, temperament/personality, interest, and biodata predictor measures was assembled and used to test several thousand soldiers as they entered in four Military Occupational Specialties (MOS). The data from this "preliminary battery sample" along with information from a large-scale literature review and a set of structured, expert judgments were then used to identify "best bet" measures. These "best bet" measures were developed, pilot tested, and refined. The refined test battery was then field tested to assess reliabilities, "fakability," practice effects, and so forth. The resulting predictor battery, now called the "Trial Battery," which includes computer-administered perceptual and psychomotor measures, will be administered together with a comprehensive set of job performance indices based on job knowledge tests, hands-on job samples, and performance rating measures in the Concurrent Validation.

In the third iteration (the Longitudinal Validation), all of the measures, refined on the basis of experience in field testing and the Concurrent Validation, will be administered in a true predictive validity design. About 50,000 soldiers across 20 MOS will be included in the FY86-87 "Experimental Predictor Battery" administration and subsequent first-tour measurement. About 3500 of these soldiers are estimated for availability for second-tour performance measurement in FY91.

For both the concurrent and longitudinal validations, the sample of MOS was specially selected as a representative sample of the Army's 250+ entry-level MOS. The selection was based on an initial clustering of MOS derived from rated similarities of job content. These MOS account for about 45 percent of Army accessions. Sample sizes are sufficient so that race and sex fairness can be empirically evaluated in most MOS.

Activities and progress during the first two years of the project were reported for FY83 in ARI Research Report 1347 and its Technical Appendix, ARI Research Note 83-37, and for FY84 in ARI Research Report 1393 and its related reports, ARI Technical Report 660 and ARI Research Note 85-14. Other publications on specific activities during those years are listed in those annual reports. The annual report on project-wide activities during FY85 is under preparation.

For administrative purposes, Project A is divided into five research tasks:

Task 1 -- Validity Analyses and Data Base Management
Task 2 -- Developing Predictors of Job Performance
Task 3 -- Developing Measures of School/Training Success
Task 4 -- Developing Measures of Army-Wide Performance
Task 5 -- Developing MOS-Specific Performance Measures

The development and revision of the wide variety of predictor and criterion measures reached the stage of extensive field testing during FY84 and the first half of FY85. These field tests resulted in the formulation of the test batteries that will be used in the comprehensive Concurrent Validation program which is being initiated in FY85.

BEST COPY AVAILABLE

The present report is one of five reports prepared under Tasks 2-5 to report the development of the measures and the results of the field tests, and to describe the measures to be used in Concurrent Validation. The five reports are:

Task 2 -- Development and Field Test of the Trial Battery for Project A, Norman G. Peterson, Editor, ARI Technical Report (in preparation).

Task 3 -- Development and Field Test of Job-Relevant Knowledge Tests for Selected MOS, by Robert H. Davis, et al., ARI Technical Report (in preparation).

Task 4 -- Development and Field Test of Army-Wide Rating Scales and the Rater Orientation and Training Program, by Elaine D. Pulakos and Walter C. Borman, Editors, ARI Technical Report 716.

Task 5 -- Development and Field Test of Task-Based MOS-Specific Criterion Measures, Charlotte H. Campbell, et al., ARI Technical Report 717.

-- Development and Field Test of Behaviorally Anchored Rating Scales for Nine MOS, Jody L. Toquam, et al., ARI Technical Report (in preparation).

BEST COPY AVAILABLE

CHAPTER 1: RATIONALE FOR THE ARMY-WIDE RATING SCALES

Walter C. Borman

Introduction

This report is divided into two major sections. The first section, which includes Chapters 1 through 4, describes the developmental procedures for the Army-wide rating scales. Specifically, Chapter 1 discusses our rationale for the Army-wide rating scales, Chapter 2 describes the development of the rating scales, and Chapter 3 presents the procedures used to field test the rating instruments. Then, Chapter 4 describes results of the field tests and our final revisions to the Army-wide scales in preparation for Concurrent Validation.

The second major section of this report is focused on rater orientation and training. Chapter 5 presents our general approach to rater training, a detailed description of the program itself, and how the program was revised based on our field test experience. Finally, Chapter 6 contains the results of an experiment conducted during the second group of field tests to evaluate particular components of the rater training and orientation program.

As mentioned, this chapter presents the rationale for Army-wide scale development, as part of the Project A effort to evaluate the validity of current and new predictors of first-term soldier job performance. A primary goal of the project is to increase Army organizational effectiveness by improving the soldier-job match. This goal will be achieved by first developing a comprehensive set of selection and classification measures (predictors) and job performance criteria, and then empirically investigating relationships between these predictor and job performance measures. Thus, one very important part of the project is to develop and utilize reliable and valid measures of job performance.

The Project A research is being conducted on 19 Army jobs (Military Occupational Specialties or MOS), carefully selected to be representative of the entire population of Army MOS. Accordingly, performance criterion measures are required for each of these 19 MOS.

This criterion development effort is well under way. Hands-on, task proficiency performance tests (Campbell et al., 1985) and job knowledge tests (Davis et al., 1985) have been developed and field tested for nine of the jobs. In addition, behavior-based rating scales have undergone similar development and field testing procedures, again for each of the nine jobs (Toquam et al., 1985). It should be noted that these diverse and comprehensive criterion development activities are a cornerstone of the Project A work. Considerable effort is being extended to ensure that the criterion measures reflect all important performance requirements of the MOS involved in this research.

BEST COPY AVAILABLE

Time and cost limitations dictated that criterion measurement coverage be provided for just nine of the 19 target jobs with the job-specific hands-on, job knowledge, and rating scale measures. Yet, it was seen as critically important to have performance criterion measures for the other jobs in this representative sample of MOS. Also, for future research efforts requiring criterion measurement on a wider sampling of MOS, it seemed desirable to have a performance measure relevant to all first-term Army jobs.

In response to these needs, the Army Research Institute in the Statement of Work for Project A requested the development of performance rating scales that could be used to evaluate soldier effectiveness in any MOS. The present scale development effort directly addresses this requirement.[1] The research objective was to prepare a single set of behavior-based rating scales relevant to all first-term soldiers.

Another objective was to develop rating scales that focused on tasks that all first-term soldiers are required to perform. As will be explained later in the report, the Skill Level I Common Task Soldier's Manual provided an excellent source for task areas to include on these task rating scales.

A Model of Soldier Effectiveness

As part of the rationale for the Army-wide rating scales, we developed a conceptual model of first-term soldier effectiveness (Borman, Motowidlo, Rose, and Hanser, 1984). In this model, we sought to expand the set of criterion behaviors considered to include elements of individual effectiveness not directly related to task performance, but related instead to a broader conception of job performance factors.

In particular, elements were considered if they appeared to be potentially important contributors to organizational effectiveness in Army units. The notion here was that being a good soldier from the Army's perspective means more than doing the job properly in terms of performing tasks in a technically proficient manner. With this framework, a model of soldier effectiveness may include elements apart from MOS job performances if they contribute to a soldier's effectiveness in the unit and to his or her "overall worth to the Army."

---

[1] Two other efforts in Project A are also directed toward measuring performance in MOS other than those nine mentioned above. School job knowledge tests were developed for each of the other 10 jobs in the sample of 19 (Davis, et al., 1985), and objective performance indicators from Army records have been identified to index effectiveness in any MOS (Riegelhaupt, Harris, & Sadacca, 1985).

BEST COPY AVAILABLE

The conceptual model appears in Figure 1. It is a result of preliminary hypotheses about constructs that might be considered under the broad soldier effectiveness domain (Borman, Motowidlo, & Hanser, 1983). As depicted in the model, the constructs revolve around the areas of organizational commitment, organizational socialization, and morale. See Borman et al. (1984) for a more complete description of the elements in the model.

The conceptual model was considered important to guide thinking in subsequent scale development steps. However, we also believed strongly that an empirical strategy should be used to examine the soldier effectiveness domain. Accordingly, a variant of the critical incident or behavioral analysis (Smith and Kendall, 1963) approach was employed to identify dimensions of soldier effectiveness. The adequacy and comprehensiveness of our empirically derived dimensions were then assessed by comparing them to those represented in the preliminary model of soldier effectiveness. Specific procedures are presented in Chapter 2.

BEST COPY AVAILABLE

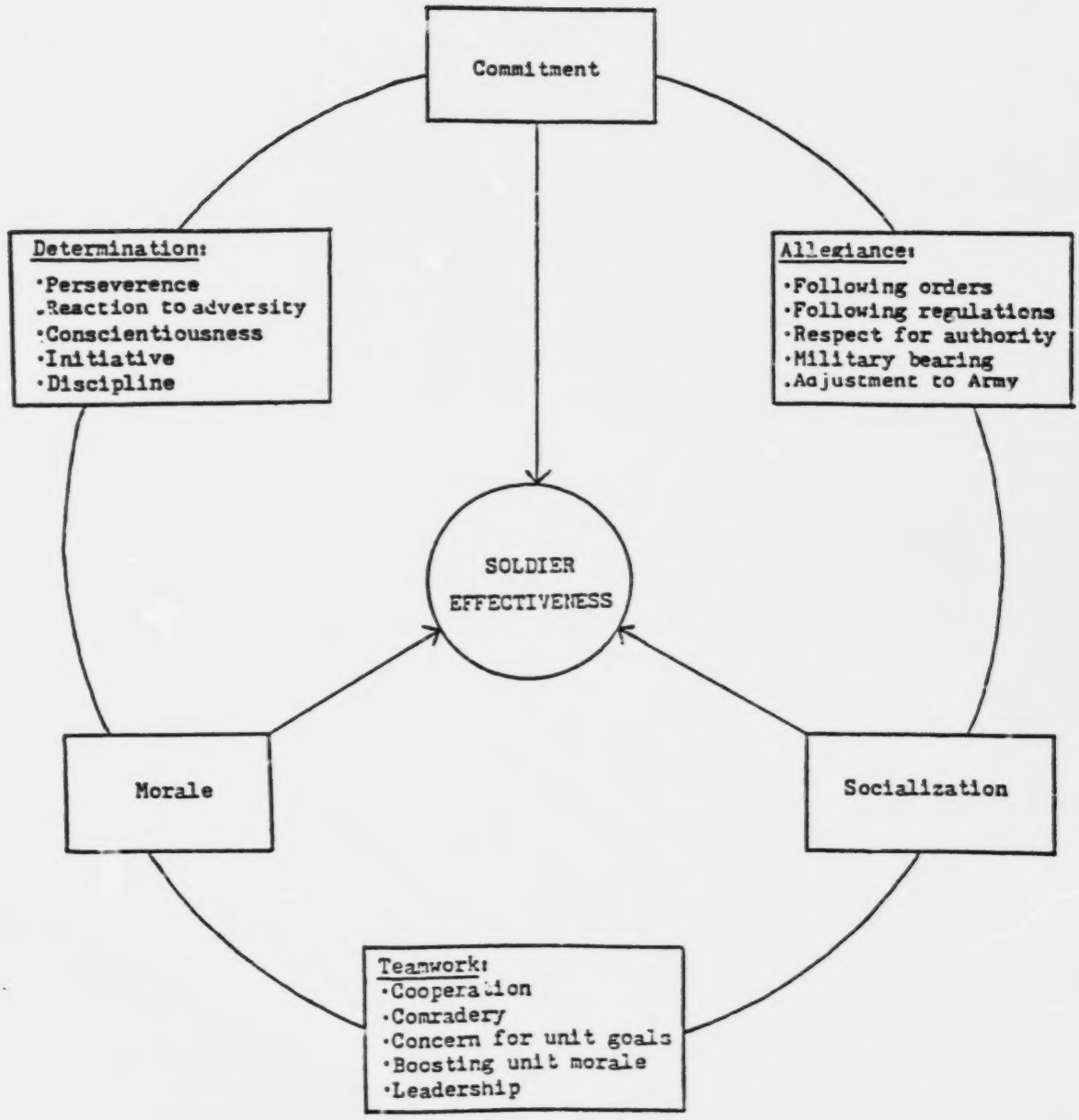


Figure 1. A Preliminary Model of Soldier Effectiveness

BEST COPY AVAILABLE

CHAPTER 2: DEVELOPMENT OF THE ARMY-WIDE RATING SCALES AND TASK DIMENSIONS

Walter C. Borman and Sharon R. Rose

Summary of Procedures

The inductive behavioral analysis strategy (Campbell, Dunnette, Arvey, & Hellervik, 1973) requires persons familiar with a job's performance demands to generate examples of effective, mid-range, and ineffective behavior observed on that job. In the present application, "job behavior" was defined broadly as any action related to soldier effectiveness, and officer and non-commissioned officer (NCO) participants in workshops were asked to generate behavioral examples from any aspect of what they considered to be the first-term soldier effectiveness domain. Behaviors generated were to be appropriate for and applicable to any MOS.

The many behavioral examples emerging from this step were first content analyzed to form dimensions or categories of soldier effectiveness and then submitted to a retranslation procedure. In retranslation, officers and NCOs evaluated each example, placing it in a category and rating the level of effectiveness reflected. Examples that showed good agreement in the retranslation step were used to form behavioral statements anchoring different levels of effectiveness on each of the dimensions. These dimensions, with their behavioral anchors, then served as the Army-wide rating scales.

In addition, Army-wide task dimensions were generated based on the tasks appearing in the Skill Level I Common Task Soldier's Manual. These dimensions were then put in a form appropriate for performance rating scales and became the Army-wide common task rating scales.

Army-Wide Rating Scales Behavior Analysis Workshops

Seventy-seven officers and NCOs participated in six one-day workshops intended primarily to elicit behavioral examples of soldier effectiveness. Table 1 describes the workshop participant groups.

In each workshop, the leader first provided an introductory briefing, describing the Project A program. The workshop leader then distributed the orientation materials that appear in Appendix A. A very important section of these materials is the training program designed to help workshop participants write appropriate behavioral examples.

Regarding this training, the workshop leader first provided general instruction on how to write soldier effectiveness incidents. Specifically, participants were told to try and remember what a soldier actually did or failed to do that made him or her effective or ineffective in a situation. Also emphasized was that participants describe only what they saw or what the soldier did, not what they

BEST COPY AVAILABLE

Table 1

Participants in Behavior Analysis Workshops

Rank	n	Sex	n
NCOs			
SP4	1	Male	28
E-5	5	Female	2
E-6	13		
E-7	11		
Officers			
First Lt.	3	Male	44
Captain	29	Female	3
Major	15		

BEST COPY AVAILABLE

inferred from the action. For example, rather than writing that a soldier "displayed loyalty," the group was told to describe precisely what the soldier did that showed he or she was loyal (e.g., worked all night to accomplish a job or spoke very highly of his or her CO). The features of a good incident were then reviewed with participants as follows:

1. It concerns the actions of an individual soldier.

2. It tells what the soldier did (or did not do) that made you feel he or she was effective or ineffective.

3. It describes clearly the background of the incident.

4. It states consequences of what the soldier did.

5. It is concise in that it is short, to the point, and does not go to great lengths specifying unimportant details of the background, the activity itself, or the consequences of what the soldier did.

The second major component of training had a modeling orientation in which participants were shown improperly written behavioral examples and, importantly, these examples corrected to their proper form. The examples focused on common errors that are made in writing critical incidents and how to avoid making these errors. The frequently encountered errors that were discussed in training included: (1) not providing sufficient information to evaluate the soldier's behavior; (2) not clearly describing what the soldier did; (3) not clearly describing the result of the soldier's action; (4) including irrelevant information; (5) labeling the behavior rather than indicating what the soldier actually did; and (6) writing "doubled-barreled" incidents (i.e., incidents that contain both positive and negative behaviors and/or consequences and are thus ambiguous in terms of their effectiveness).

Participants were led through this training and then asked to write a first behavioral example. Workshop leaders reviewed the first example and provided corrective guidance as needed. This step was important in order to ensure that each participant was writing appropriate behavioral examples of incidents. Except for periods taken to discuss behavioral examples or effectiveness dimensions emerging from the content of the examples, the remainder of each workshop was devoted to participants' writing and leaders' reviewing the examples. Below are two such examples to provide a flavor for the output from the workshops.

- This soldier was in a group sitting around a tree when a senior officer walked toward them. He called the group to attention and saluted the officer.

BEST COPY AVAILABLE

- When this soldier was assigned to guard a bivouac area at night on an FTX, he fell asleep at one of the training stations even though he knew he was supposed to be walking the post.

A total of 1,315 behavioral examples were generated in the six workshops. Details relevant to this data collection appear in Table 2. Duplicate examples and those examples which did not meet the criteria specified in training (e.g., the incident described the behavior of an NCO rather than a first-term soldier, were dropped from further consideration.

The remaining 1,111 examples were edited to a common format. The senior author supervised the editing process to ensure that uniform guidelines were followed by editors and to guard against interpretive biases entering into the editing process. All edited incidents were then content analyzed to form preliminary dimensions of soldier effectiveness. Specifically, three researchers independently read each example and grouped together those examples that described similar behaviors. The sorted examples were then reviewed and the groupings were revised, yielding a set of 13 dimensions that were homogeneous with respect to their content. These 13 dimensions were then reviewed by a small group of officers and NCOs at Fort Benning, and a consensus was reached that the proposed dimensions were suitable for further scale development work. The 13 dimensions were as follows:

A. Controlling own behavior related to personal finances, drugs/alcohol, and aggressive acts.
B. Adhering to regulations and SOP, and displaying respect for authority.
C. Displaying honesty and integrity.
D. Maintaining proper military appearance.
E. Maintaining proper physical fitness.
F. Maintaining own equipment.
G. Maintaining living and work areas to Army-unit standards.
H. Exhibiting technical knowledge and skill.
I. Showing initiative and extra effort on the job/mission/assignment.
J. Attending to detail on jobs/assignments/equipment checks.
K. Developing own job and soldiering skills.
L. Effectively leading and providing motivation to other soldiers.
M. Supporting other unit members.

BEST COPY AVAILABLE

Table 2

Soldier Effectiveness Examples Generated

Location	Participants	Number of Examples	Mean Examples Per Participant
Fort Benning	14 Officers	228	16
Fort Stewart	13 Officers	266	20
Fort Stewart	13 NCOs	216	17
Fort Knox	12 Officers	239	20
Fort Benning	13 NCOs	149	11
Fort Carson	8 Officers		
	4 NCOs	217	18
Totals:	77	1,315	17

BEST COPY AVAILABLE

Retranslation of the Behavioral Examples

Retranslation provides a way of checking on the clarity of individual behavioral examples and of the dimension system. As mentioned, in retranslation, persons familiar with the target domain make two judgments about each example: the dimension or category to which it belongs based on its content, and the effectiveness level it reflects. Examples for which there is disagreement related either to category membership or to the rated effectiveness level may be unclear and should be revised or eliminated from further consideration. Also, confusion between two or more categories in the sorting of several examples may reflect a poorly formed and/or defined category system.

In this project, the retranslation task was divided into five subtasks, each requiring a retranslation judge to evaluate 216-225 behavioral examples. The retranslation judges were a different group of individuals than those who generated the critical incidents. The division into subtasks was accomplished to keep reasonable the amount of time each judge would need to spend on the rating task. Judges were provided with definitions of each of the 13 dimensions to aid in the sorting and a 1-9 effectiveness scale (1 = extremely ineffective; 5 = adequate/average; and 9 = extremely effective) to guide the effectiveness ratings. The retranslation materials, including all 1,111 edited behavioral examples, appear in Appendix B. Sixty-one officer and NCO judges completed retranslation ratings, and these results are presented below.

Retranslation Results

Table 3 shows the number of behavioral examples reliably retranslated for each of the 13 dimensions. Typically employed acceptance points were greater than 50 percent for sorting an example into a single dimension, and less than a 2.0 standard deviation for the effectiveness ratings. This process left 870 of the 1,111 examples (78%) to be included for subsequent scale development work. Appendix B contains effectiveness scale means and standard deviations for each behavioral example, along with the percentage of retranslation raters sorting each example into each dimension.

Most of the dimensions (shown in Table 1) that were developed by the empirical procedures are quite consistent with the dimensions that were theoretically expected according to our preliminary conceptual model. Empirical dimensions A, I, J, and F seem to capture elements of the "Determination" category in our model. Dimensions B, D, and E reflect elements of the "Allegiance" category, and dimensions L and M reflect "Teamwork." This convergence with the theoretically expected dimensions gives us confidence that the empirically derived dimensions tap important, generalizable facets of soldier effectiveness.

BEST COPY AVAILABLE

Table 3

Number of Behavioral Examples Reliably Retranslated Into Each Dimension

Dimension	Number of Examples
A. Controlling own behavior related to personal finances, drugs/alcohol, and aggressive acts	107
B. Adhering to regulations and SOP, and displaying respect for authority	158
C. Displaying honesty and integrity	53
D. Maintaining proper military appearance	34
E. Maintaining proper physical fitness	36
F. Maintaining own equipment[a]	46
G. Maintaining living and work areas to Army-unit standards	23
H. Exhibiting technical knowledge and skill	47
I. Showing initiative and extra effort on job/ mission/assignment	131
J. Attending to detail on jobs/assignments/ equipment checks[a]	59
K. Developing own job and soldiering skills	40
L. Effectively leading and providing motivation to other soldiers[b]	71
M. Supporting other unit members[b]	65
	870

Note. Examples were retained if they were sorted into a single dimension by greater than 50% of the retranslation raters and the standard deviations of their effectiveness ratings were less than 2.0.

[a] These two dimensions were subsequently combined to form a Leadership dimension.

[b] These two dimensions were subsequently combined to form a Maintaining Assigned Equipment dimension.

BEST COPY AVAILABLE

Results in Table 3 were seen as satisfactory in that sufficient numbers of reliably retranslated examples were available to develop behavioral definitions of each dimension. Typically, a minimum of 20 reliably retranslated examples that are not highly overlapping in content is considered to be sufficient for defining a dimension. However, in the spirit of shortening the rating task for subsequent rating scale administrations, two pairs of dimensions were combined. Specifically, Leading Other Soldiers and Supporting Other Unit Members were combined to form Leading/Supporting, and Attending to Detail and Maintaining Own Equipment were collapsed to form a single dimension titled Maintaining Assigned Equipment. These two collapsings, resulting in a total of 11 Army-wide dimensions, were deemed justifiable because of the conceptual similarity of each of the dimension pairs.

For each dimension, the reliably retranslated behavioral examples were divided into three categories of effectiveness levels: low (1-3.49), average (3.5-6.49), and high (6.5-9). Behavioral summary statements were then written to capture the content of the specific examples at these three performance levels.

Development of the behavioral summary statements is the critical step in forming Behavior Summary Scales. The main advantage of these scales over behaviorally anchored rating scales is that, for a particular dimension and level of effectiveness, the content of the examples reliably retranslated is represented on the scale, not just one or two of the specific behavioral examples (Borman, 1979). Accordingly, it is more likely that a rater using the scales will be able to match observed performance with the performance descriptions that appear on the scales. The 11 behavior-based rating scales were thus readied for administration in the field tests. These scales appear in Appendix C.

In addition, two summary rating scales were prepared. First, an overall effectiveness scale was developed to obtain overall judgments of a soldier's effectiveness based on all of the behavioral dimension ratings. Second, an NCO potential scale was developed to assess each soldier's likelihood of being an effective supervisor as an NCO. These scales also appear in Appendix C.

Development of the Army-Wide Common Task Dimensions

Rating scales covering the common task domain were developed from tasks appearing in the Skill Level I Common Task Soldier's Manual. Because this manual contains tasks that all first-term soldiers are expected to perform, it seemed to be a good source for Army-wide task dimensions. To develop these dimensions, the specific tasks contained in the manual (e.g., "Read and Report Total Radiation Dose" and "Repair Field Wire") were content analyzed, and 13 common task areas that appeared to reflect in summary form all of the specific tasks were identified. Common task areas included the following:

BEST COPY AVAILABLE

A. See: Identifying Threat (armored vehicles, aircraft)
B. See: Estimating Range
C. Communicate: Send a Radio Message
D. Navigate: Using a Map
E. Navigate: Navigating in the Field
F. Shoot: Performing Operator Maintenance on Weapon (e.g., M16 rifle)
G. Shoot: Engaging Target with Weapon (e.g., M16)
H. Combat Techniques: Moving Under Direct Fire
I. Combat Techniques: Clearing Fields of Fire
J. Combat Techniques: Camouflaging Self and Equipment
K. Survive: Protecting Against NBC Attack
L. Survive: Performing First Aid on Self and Other Casualties
M. Survive: Knowing and Applying the Customs and Laws of War

Ratings for each common task area involved evaluating on a 7-point scale how effectively the ratee typically performed. The scale ranged from 1 = "Poor: does not meet standards and expectations for adequate performance in this task area" to 7 = "Excellent: exceeds standards and expectations for performance in this task area." In addition, raters were given the option of choosing a "0," indicating that they had not observed a soldier performing in the task area. The actual rating scales appear in Appendix C.[2]

---

[2] The scales that appear in Appendix C are the version that was used for the Batch B field tests. However, some revisions were made between the Batch A and Batch B field tests (see Chapter 4 for details), and thus, some of the dimensions listed above do not correspond exactly to the dimensions appearing in Appendix C.

BEST COPY AVAILABLE

CHAPTER 3: FIELD TEST PROCEDURES FOR ADMINISTERING THE ARMY-WIDE RATING SCALES

Elaine D. Pulakos and Walter C. Borman

Summary of Procedures

Field tests for the MOS in Batch A were conducted during the summer and fall of 1984. Soldiers from the following four MOS participated: Administrative Specialist (71L); Motor Transport Operator (64C); Military Police (95B); and Cannon Crewman (13B). The field tests for the MOS in Batch B took place in the winter and spring of 1985 and included five MOS: Infantryman (11B); Armor Crewman (19E); Single Channel Radio Operator (31C); Light Wheeled Vehicle Mechanic (63B); and Medical Specialist (91A). Supervisor and peer ratings were gathered for 1,369 first-term soldiers (approximately 150 per MOS). See Tables 4 and 5 for a description of the total sample tested and rated.

Our goal was to have each soldier's first- and second-level supervisors provide for-research-only performance evaluations using the Army-wide scales. In addition, we sought four fellow first-term soldiers or peers who were sufficiently familiar with each subject's performance to evaluate him or her on the same scales.

Sample and Procedures

As mentioned above, our goal was to obtain ratings of each first-term soldier in the sample from two supervisors and four peers who were very familiar with the ratee's job performance. For supervisors, raters were assigned participating first-term soldiers whom they had supervised for at least two months and whose performance they knew well.

The peer rating assignments proceeded as follows. At the first meeting of a first-term soldier group going through testing, potential peer raters received an alphabetized list of all of the individuals within his or her MOS from whom data were being collected at that post. The soldiers were asked to check off the names of each peer they had worked with for at least two months and whose job performance they knew well. This procedure resulted in some soldiers checking no names (if they were the only participant from their unit) and others checking as many as 20 or more names (if many soldiers from the same unit were participating in the field test).

If an individual soldier did not check any names or checked only one or two names, he or she was questioned to ascertain why this was the case. In order to obtain a sufficient number of peer ratings for each soldier, the minimum criterion of working with a potential ratee for

BEST COPY AVAILABLE

Table 4

Number of Soldiers in the Field Tests by MOS and Location

	MOS									
Location	11B	13B	19E	31C	63B	64C	71L	91A	95B	Total
Fort Hood	--	--	--	--	--	--	48	--	42	90
Fort Lewis	29	--	30	16	13	--	--	24	--	112
Fort Polk	30	--	31	26	26	--	60	30	42	245
Fort Riley	30	--	24	26	29	--	21	34	30	194
Fort Stewart	31	--	30	23	27	--	--	21	--	132
USAREUR	58	150	57	57	61	155	--	58	--	596
Total	178	150	172	148	156	155	129	167	114	1369

BEST COPY AVAILABLE

Table 5

Soldiers in the Field Tests by Sex and Race

	Sex		
Race	Male	Female	Total
Black	330	58	388
Hispanic	37	3	40
White	789	104	893
Other	43	5	48
Total	1199	170	1369

BEST COPY AVAILABLE

at least two months was, under certain circumstances, relaxed to a minimum of at least one month. However, relaxing the criterion was done on an individual basis by administrators after they had established that the soldier in question had enough familiarity with potential ratees to evaluate their performance. Specific guidelines were provided to administrators to help them determine the circumstances under which it would and would not be appropriate to relax the two month criterion. These guidelines as well as specific administrator instructions for identifying potential peer raters are provided in Appendix D.

A computerized random assignment procedure was utilized to assign ratees to peer raters with the constraints that (1) raters were assigned only soldiers whose names they checked off; (2) ratees with few potential raters were assigned to raters early in the allocation process; and (3) the number of ratees per rater was equalized as much as possible.

The actual numbers of raters per ratee are shown in Table 6 by MOS and rater source (i.e., supervisors and peers). In all MOS, we fell short of our goal of two supervisor and four peer raters. However, with the exception of the 71L supervisors, the numbers of raters per ratee were sufficient to allow reasonable interrater reliability estimates.

The Army-wide rating scales (along with other rating measures) were administered to groups of peer or supervisor raters. The data analyses relevant to these scales are presented in the following chapter. For the peer rating sessions, the groups were typically 12-15 in size. For the supervisor sessions, anywhere from 2 or 3 to 15-20 attended.

An extremely important aspect of each rating session was the rater orientation and training program that was designed to reduce various rating errors and to persuade raters to try hard to provide accurate, for-research-only performance evaluations. This program, as well as the revisions that were made to it as a result of the field tests, is described in Chapter 5 of this report.

BEST COPY AVAILABLE

Table 6

Rater and Ratee Sample Sizes in the Field Tests

Rater Group	Number of Ratees	Number of Raters	Number of Rater/Ratee Pairs	Raters/ Ratee Ratio
Supervisors				
11B	149	126	269	1.81
13B	146	50	215	1.47
19E	161	145	270	1.68
31C	130	118	225	1.73
63B	141	145	250	1.77
64C	150	108	276	1.84
71L	109	82	114	1.05
91A	156	86	248	1.59
95B	113	44	219	1.94
Peers				
11B	172	170	515	2.99
13B	141	141	430	2.89
19E	163	156	481	2.95
31C	122	127	304	2.49
63B	129	133	268	2.08
64C	155	150	430	2.77
71L	64	61	123	1.92
91A	157	155	486	3.10
95B	113	113	415	3.67

BEST COPY AVAILABLE

CHAPTER 4: ARMY-WIDE RATING DATA ANALYSES AND RESULTS

Walter C. Borman and Elaine D. Pulakos

Data Analyses

For each MOS, ratings collected during the field tests were analyzed across posts. These analyses were conducted separately for the peers and supervisors. For each of these rater groups, analyses of the Army-wide data focused on (1) distributions of the ratings, (2) intercorrelations of the rating scale dimensions, (3) interrater reliabilities, and (4) relationships between the Army-wide rating scales and other rating measures.

Prior to conducting these analyses, however, inspection of the rating data revealed level differences in the mean ratings provided by two or more raters of the same ratees. Since our primary interest was each soldier's profile of ratings across the different behavioral and common task dimensions (i.e., ratees' relative strengths and weaknesses compared to each other), all raters' scores were adjusted to eliminate any level differences between them.

Outlined below are the procedures that were used to compute the adjusted ratings, including an example for one rater to demonstrate these procedures. These analyses were conducted separately for the behavioral rating dimensions (including Overall Effectiveness and NCO Potential) and for the common task dimensions.

- For each rater, a mean score was computed across dimensions for each soldier ratee. If, for example, Rater 1 evaluated two soldiers on the behavioral dimensions, two means were calculated.

	Ratee Means Across All Behavioral Dimensions
Rater 1	
Soldier A	3.30
Soldier B	4.50

- All other peer and supervisor raters providing evaluations for the same two target soldiers were then identified. For each soldier, a mean rating was calculated across all raters and dimensions. In our example, assume that Raters 2, 3, and 4 evaluated Soldier A. The mean was computed for Soldier A across the behavioral dimensions and across the three raters; this mean was 4.2. Assume that Rater 3 also evaluated Soldier B and that he was the only other rater for that soldier. We calculated Rater 3's mean rating of Soldier B across the behavioral dimensions; this mean was 4.3

BEST COPY AVAILABLE

	Ratee Means Across All Behavioral Dimensions	Ratee Mean Across All Behavioral Dimensions & Raters
Soldier A		
Rater 2	4.00	
Rater 3	4.40	4.20
Rater 4	4.20	
Soldier B		
Rater 3	4.30	4.30

- An individual rater's mean across dimensions for a given soldier was then compared with the mean rating computed for the same soldier across all other raters and dimensions. A difference score was computed between these two mean values.

(Rater 1's Mean for Soldier A)	-	(All Other Raters' Mean for Soldier A)	= Difference Score
3.30	-	4.20	= -0.9

(Rater 1's Mean for Soldier B)	-	(All Other Raters' Mean for Soldier B)	= Difference Score
4.50	-	4.30	= 0.2

  Rater 1 would thus receive a difference score of -0.9 for Soldier A and a difference score of 0.2 for Soldier B. This procedure was repeated, computing a difference score between each rater's mean rating for each of his or her ratees and the mean rating provided by all other raters of the same ratees.

- Next, the rater's difference score for each ratee was weighted by the total number of other raters who evaluated that target ratee. In the present example, Rater 1 had a difference score of -0.9 for Soldier A and 0.2 for Soldier B. The difference score for Soldier A was weighted 3 because three other raters evaluated the soldier, while the difference score for Soldier B was weighted 1 because only one other rater evaluated the soldier.

- For each rater, an average weighted difference score was then computed and used to adjust the ratings for all soldiers evaluated by that rater.

BEST COPY AVAILABLE

For Rater 1, the weighted average difference score was as follows:

$$[(3(-0.9) + (0.2))/4] = -.625$$

Thus, Rater 1's ratings on each dimension for Soldiers A and B were increased by a value of 0.625.

These procedures were used to compute adjusted scores for all raters. Recall that these adjustments were performed separately for the behavioral dimensions and the common task dimensions. All of the analyses presented in this report were conducted using the adjusted rating data.

Identification of Outlier Raters

The rating data were first inspected for the purpose of identifying potential outlier raters. That is, we were interested in pinpointing individual raters whose ratings might be so severely discrepant from other raters' ratings of the same ratees that the validity of these ratings was suspect. Two criteria were used to identify such raters:

- First, correlations were computed across all rating instruments[3] between each individual rater's ratings of all his or her ratees and the average of all other raters' ratings of these same ratees. If this correlation was less than -.20, the individual rater was identified an outlier.

- Second, if across all of the rating instruments, an individual rater's mean adjustment was greater than ± 2.5 (on the 7-point scales) the rater was identified as an outlier.

Using these two criteria, nine supervisors and 46 peers were identified as outliers and excluded from further analyses.

---

[3]The criteria for identifying outlier raters were based on data from all of the rating instruments administered during the field tests. These included the Army-wide behavioral dimensions and common task dimensions described in this report, MOS-specific behavioral dimensions (Toquam, et al., 1985), MOS-specific task dimensions (Campbell et al., 1985), ratee characteristic dimensions (Borman, White, and Gast, 1984), and combat performance prediction dimensions (Reigelhaupt et al., 1985).

BEST COPY AVAILABLE

Results

Distributions of Ratings

One criterion for assessing the quality of the supervisor and peer ratings is to evaluate the distributions of those ratings. Particularly in operational settings, ratings tend to be very skewed, with most ratees receiving high performance ratings. For-research-only administrations of rating scales (such as the present effort) often yield ratings that are more normally distributed, with lower mean ratings and greater variance in evaluations across ratees. However, since there is always concern about non-normal distributions of ratings, we examined this element first.

Table 7 presents frequency distributions of ratings made on each of the seven points on the 7-point rating scales. Table 8 then depicts the means, and standard deviations of selected composite ratings as well as the Overall Performance and NCO Potential scales.

Taken together, findings from the two tables suggest that raters did not succumb to excessive leniency (overly high ratings) or restriction-in-range (rating everyone at about the same level). The modal rating of 5 on a 7-point scale and means generally between 4 and 5 seem reasonable in that we would expect the average performance of a first-term soldier be a little above average. The rationale underlying this assumption is that some percentage of the poor performers will have already left the Army. Likewise, the spread of the ratings across the seven scale points seems reasonable. Although the lowest point (1) was used in relatively few cases, the frequencies with which raters employed the other scale points were at the expected levels. In addition, the standard deviations of the ratings suggest good spread.

Within-Instrument Intercorrelations

Appendix E contains the correlations between the ratings on the Army-wide behavioral dimensions, and Appendix F shows the corresponding correlations for the common task scale dimensions. The correlations were quite high within both rating instruments for all MOS, suggesting that raters did not differentiate very much between dimensions when making their evaluations on these scales. Accordingly, a unit weighted composite of the dimension ratings was computed for each of the behavioral and the common task scales. These composites were used in performing most of the subsequent analyses.

Interrater Reliability

Interrater reliabilities for selected Army-wide measures appear in Table 9. In general, the reliabilities were encouraging. Intraclass correlation coefficients (ICCs) for the composites of the Army-wide behavioral dimensions were almost uniformly in the .80s (mdn = .84). Reliabilities of the individual behavioral scales were lower (.46-.68, mdn = .58) but still very respectable. The Overall Effectiveness and

BEST COPY AVAILABLE

Table 7

Frequency Distributions of Ratings Across the Seven Rating Scale Points (In Percents)

	Scale Points						
MOS	1	2	3	4	5	6	7
Individual Army-wide Behavioral Dimensions							
11B	4/2	8/6	13/13	20/23	25/31	18/18	11/8
13B	3/4	7/6	13/11	17/18	24/30	23/23	13/9
19E	2/2	8/7	13/15	22/24	28/30	18/15	10/6
31C	3/3	9/6	14/12	19/18	30/32	16/19	9/10
63B	2/2	10/7	16/13	20/21	27/31	15/17	9/9
64C	4/3	9/6	12/12	19/21	26/34	18/18	12/6
71L	2/1	6/4	14/12	17/25	26/30	20/19	15/8
91A	4/3	9/8	12/13	19/22	27/28	18/18	12/9
95B	2/2	6/6	15/16	25/26	29/30	16/17	8/4
Overall Effectiveness Dimension							
11B	2/1	6/2	17/12	24/26	30/37	16/19	4/3
13B	2/3	6/5	16/8	25/15	24/36	17/26	10/8
19E	0/1	4/4	10/12	25/28	42/33	15/19	3/3
31C	1/1	6/4	15/8	25/21	36/37	12/21	4/8
63B	2/1	8/4	14/12	29/25	30/38	14/15	4/4
64C	3/2	9/5	18/7	18/16	30/42	17/24	6/3
71L	0/0	7/3	15/9	27/31	29/32	20/23	2/2
91A	2/1	5/4	16/13	24/26	27/35	21/17	5/4
95B	1/1	4/3	14/11	23/27	32/38	20/19	7/2
NCO Potential Dimension							
11B	9/4	18/12	15/18	17/19	22/25	15/17	5/4
13B	11/7	5/7	10/8	18/12	27/32	17/21	10/13
19E	3/6	11/10	13/17	21/24	27/26	20/13	5/4
31C	7/5	15/6	14/10	16/18	27/25	15/27	5/9
63B	6/4	13/10	17/14	21/19	21/29	15/17	6/7
64C	8/8	6/10	13/13	21/20	27/30	16/15	8/4
71L	2/0	4/3	11/12	18/24	38/39	19/15	10/7
91A	8/7	14/10	14/15	15/20	22/25	18/16	9/8
95B	6/7	5/7	10/15	21/23	25/24	18/18	14/7
Individual Army-wide Common Task Dimensions							
11B	2/1	6/4	13/10	18/20	27/29	20/23	15/13
13B	3/3	5/3	9/8	19/15	28/28	23/29	13/14
19E	1/1	3/3	9/10	19/22	33/28	22/24	13/13
31C	1/1	3/3	9/7	20/19	29/30	22/24	16/16
63B	2/2	5/3	12/12	20/20	30/29	21/23	10/12
64C	3/3	5/6	12/11	20/20	28/34	25/19	7/6
71L	2/3	6/6	12/15	20/19	25/31	26/23	8/4
91A	2/3	4/4	10/10	18/20	27/27	25/21	16/14
95B	0/2	2/4	11/12	26/25	31/32	20/20	10/4

Note. In each case, the percent for supervisors is on the left and the percent for peers is on the right.

BEST COPY AVAILABLE

Table 8

Means and Standard Deviations (SD) of Selected Army-Wide Measures

	11B	13B	19E	31C	63B	64C	71L	91A	95B
Mean & SD: Ave. Behavioral Dimensions									
Supervisors	4.50 (.82)	4.76 (.90)	4.46 (.65)	4.59 (.88)	4.42 (.87)	4.52 (.91)	4.73 (.77)	4.56 (.95)	4.44 (.79)
Peers	4.53 (.68)	4.67 (.75)	4.47 (.59)	4.54 (.76)	4.49 (.76)	4.56 (.71)	4.78 (.65)	4.57 (.81)	4.46 (.66)
Mean & SD: Overall Effectiveness									
Supervisors	4.47 (1.02)	4.59 (1.23)	4.48 (.94)	4.55 (1.08)	4.38 (1.14)	4.36 (1.22)	4.39 (1.16)	4.57 (1.18)	4.56 (1.10)
Peers	4.62 (.76)	4.85 (.99)	4.67 (.76)	4.71 (.95)	4.52 (.95)	4.75 (.95)	4.73 (.93)	4.63 (.93)	4.64 (.91)
Mean & SD: NCO Potential									
Supervisors	3.97 (1.37)	4.34 (1.55)	4.26 (1.23)	4.28 (1.42)	4.14 (1.36)	4.30 (1.37)	4.76 (1.27)	4.23 (1.48)	4.59 (1.35)
Peers	4.14 (1.08)	4.66 (1.27)	4.23 (1.06)	4.56 (1.24)	4.31 (1.18)	4.14 (1.26)	4.76 (.93)	4.29 (1.27)	4.35 (1.13)
Mean & SD: Ave. Common Task Dimensions									
Supervisors	4.87 (.66)	4.97 (.70)	5.02 (.55)	5.07 (.61)	4.87 (.65)	4.53 (.63)	4.53 (.81)	4.91 (.71)	4.70 (.53)
Peers	4.96 (.61)	4.99 (.68)	4.93 (.47)	5.12 (.61)	4.84 (.77)	4.54 (.56)	4.75 (.69)	4.95 (.68)	4.63 (.57)

Note. The means, standard deviations, interrater reliabilities, and intercorrelations for each individual Army-wide behavioral dimension appear in Appendix E. The means, standard deviations, interrater reliabilities, and intercorrelations for each individual Army-wide common task dimension appear in Appendix F.

BEST COPY AVAILABLE

Table 9

Intraclass Correlation Coefficients for Selected Army-Wide Measures

	11B	13B	19E	31C	63B	64C	71L[a]	91A	95B
				ICCs for Ave. Behavioral Dimensions					
Supervisors	82	81	86	83	84	84	--	81	85
Peers	80	83	78	86	84	85	82	86	88
				Mean ICCs across Individual Behavioral Dimensions					
Supervisors	58	58	46	60	60	58	--	60	63
Peers	55	61	55	60	57	58	51	67	68
				ICCs for Overall Effectiveness					
Supervisors	64	62	54	70	63	72	--	74	82
Peers	47	60	48	65	71	66	70	68	79
				ICCs for NCO Potential					
Supervisors	74	61	53	71	63	68	--	64	68
Peers	57	63	59	74	66	69	60	69	68
				ICCs for Ave. Common Tasks					
Supervisors	77	70	74	55	55	60	--	71	74
Peers	78	72	67	64	84	65	57	79	82
				Mean ICCs across Individual Common Tasks					
Supervisors	42	48	38	38	42	--	--	46	41
Peers	51	47	46	41	51	34	33	60	57

[a]ICCs were not computed for 71L supervisor raters because almost all of the ratees were evaluated by only one supervisor.

BEST COPY AVAILABLE

NCO Potential reliabilities were likewise reasonably high (.47-.82, mdn = .66). Regarding the Army-wide common task ratings, interrater reliabilities for the dimension composites were satisfactory (.55-.84, mdn = .71), but not as high as the behavioral dimension composites. Individual common task scale interrater reliabilities were lower (.33 - .60, mdn = .44).

Supervisor and peer ratings had very similar levels of interrater reliability. For all of the measures presented in Table 9, median reliabilities were computed for supervisors and peers separately. Overall, the peer ratings were slightly higher in average reliability than those of the supervisors (supervisors: behavioral dimension mdn = .66, task mdn = .55; peers: behavioral dimension mdn = .68, task mdn = .59).[4]

Relationships Between Army-Wide Ratings and Other Rating Instruments

Table 10 presents correlations between Army-wide ratings and the job-specific rating scales measures (see Toquam et al., 1985 for a complete description of the MOS-specific behavioral scales and Campbell et al., 1985 for a complete description of the MOS-specific task rating scales). The table summarizes these relationships, averaged across MOS. Correlations between the rating measures for individual MOS appear in Appendix G.

---

[4]It should be noted that the data in Table 9 are intraclass correlation coefficients representing the reliabilities of mean ratings across supervisors or peers and, accordingly, are dependent to a degree on the average number of raters per ratee. Larger rater/ratee ratios yield higher reliabilities as a function of the Spearman-Brown Formula (similar to adding items to a test for increasing its reliability). Considering the present rater/ratee ratios (about 2.8 for peers versus 1.8 for supervisors), it is likely that the supervisor ratings would have been somewhat more reliable than peer ratings if each source had had the same number of raters per ratee. However, the coefficients appearing in the table provide the appropriate reliability estimates (of the mean supervisor and mean peer ratings), because correlations between the rating data and other variables were calculated using the mean supervisor and mean peer rating for each ratee. That is, ratings of a given ratee were averaged across supervisors and across peers, and all of the correlations reported here were computed on these means. Thus, the sample size for each correlation is the number of ratees on which it was calculated.

BEST COPY AVAILABLE

Table 10

Correlations Between Rating Measures Averaged Across All MOS

	1	2	3	4	5	6	7
1. Ave. A-W Behavioral Dimensions	--	82	79	76	66	54	57
2. Overall Effectiveness	82	--	75	70	63	52	55
3. NCO Potential	76	71	--	64	58	48	52
4. Ave. MOS Behavioral Dimensions	71	64	60	--	81	58	66
5. Overall MOS Job Performance	69	67	59	74	--	54	60
6. Ave. A-W Common Tasks	60	57	50	59	52	--	52
7. Ave. MOS Tasks	62	58	54	70	54	58	--

Note. Correlations above diagonal are for peers; correlations below diagonal are for supervisors.

BEST COPY AVAILABLE

As can be seen, correlations between rating measures are almost uniformly high. For supervisor ratings, the range is .50-.82 (mdn = .60); for peer ratings the range is .49-.82 (mdn = .60). Just as raters made limited distinctions between dimensions within-instrument, they apparently differentiated very little between MOS-specific job and task performance, performance on the common tasks, and Army-wide soldier effectiveness.

It is difficult to evaluate the correlations in Table 10. On the one hand, lower correlations across the different instruments might be expected. Scale development work showed a definite conceptual distinction between MOS-specific job/task performance and total effectiveness as a soldier. The Army-wide behavioral dimensions include the job performance component of effectiveness but add such elements as leading and supporting, self-development, and military appearance. High correlations across these sets of scales might indicate a failure on the part of most raters to make the proper distinctions between these components. On the other hand, the relatively high across-rating instrument correlations may reflect valid measurement of substantially related aspects of job performance.

Revision of the Army-Wide Scales

Experience administering the Army-wide rating scales during the Batch A field tests indicated that certain soldiers in some of the MOS had difficulty with the amount of reading required in completing the ratings. It thus seemed prudent to reduce the length of the behavioral anchors on the Army-wide behavior-based scales for the Batch B field tests and Concurrent Validation. This was accomplished by editing each behavioral statement to remove unnecessary language and reduce the reading difficulty without, however, changing the effectiveness level or meaning of the anchor itself.

In addition, it was felt that a few of the statements anchoring the different effectiveness levels were multidimensional. That is, the example behaviors contained in certain individual anchors were sufficiently different to cause raters potential confusion regarding the level at which a ratee should be evaluated. This potential problem was addressed by extrapolating more global performance information from the specific behaviors and writing the scale anchors to reflect these more general performance levels.

A second revision between the Batch A and Batch B administrations was to drop one of the 13 common task scales. This was done simply because a 13th scale would have required an additional page on the printed version of the scales. The task dimension that had the lowest interrater reliability (i.e., Combat Techniques: Clearing Fields of Fire) and seemed the most redundant with others was eliminated.

BEST COPY AVAILABLE

After the Batch B administration, the instruments were submitted to proponent review. In this review, technical school subject matter experts studied the scales and made suggestions for minor wording changes on some of the anchors. Also, the dimension Maintaining Living/Work Areas was dropped to further reduce the length of time required to complete the behavioral rating scales. Proponent review experts judged that dimension to be the least important and the most expendable. Finally, because of low interrater reliabilities, a second common task scale (i.e., Navigate: Using a Map) was dropped subsequent to the Batch B field tests.

In summary, only minimal changes were made to the Army-wide rating scales as a result of the field tests. These included eliminating one of the behavioral dimensions and two of the common task dimensions. In addition, relatively minor wording changes and reductions to the length of the scale anchors were made to lessen the reading difficulty as well as the total amount of time required to complete the scales. The Army-wide rating scales that appear in Appendix H represent the version as revised for Concurrent Validation.

BEST COPY AVAILABLE

CHAPTER 5: RATER ORIENTATION AND TRAINING

Elaine D. Pulakos and Walter C. Borman

Overview of the Approach

In the present project, rater orientation and training were considered in very broad terms. The general intent was to do everything within our power to obtain accurate ratings of individual soldier effectiveness. To accomplish this objective, a three-part approach was employed.

First, criteria were carefully laid out for selecting raters to participate in the research. As mentioned previously, to provide ratings of each soldier's effectivness on the job, we sought two supervisors and four first-tour soldiers who were familiar with the ratee's performance.

The second part of the approach related to the actual rating sessions. In one sense, these sessions can be viewed as persuasive presentations. Our intent was to persuade participants to help us in the research effort by trying hard to make accurate performance ratings. Elements of the "sale" included, (1) convincing raters that the ratings would be kept confidential and used for research purposes only, and (2) motivating raters to take the rating task seriously and to do their best to provide valid performance evaluations. The first part of the briefing was designed specifically to accomplish the convincing and motivating requirements for the rating sessions.

The third aspect of our orientation and training approach also involved the rating sessions. Besides convincing participants that the project and their role in it were legitimate and worthwhile, it was necessary to train raters to avoid certain common rating errors and to be as accurate as possible in their evaluations. This part of the strategy was especially critical for the peer rating groups. These soldiers typically have no experience in making performance evaluations and thus required special guidance in how to make the ratings and avoid the rating errors. Supervisor raters often have some experience using the Enlisted Evaluation Report (EER). However, the behavior-based scales used in the present research were more complicated than the EER and therefore required focused rater training. Accordingly, a rater training component was carefully developed to help supervisor and peer raters reduce certain rating errors and provide relatively accurate evaluations.

The rater orientation and training program was seen as very important for reaching the objective of obtaining high quality ratings. Recent reviews of research on rater training conclude that training is likely to improve performance appraisals (Landy & Farr, 1980; Zedeck & Cascio, 1982). Research has shown that rating errors such as halo and leniency can be reduced by appropriate training (Borman, 1975, 1979;

BEST COPY AVAILABLE

Brown, 1968; Latham, Wexley, & Pursell, 1975). Also, the accuracy of performance ratings has been enhanced using rater training programs (McIntyre, Smith, & Hassett, 1984; Pulakos, 1984, 1986).

Our own experience suggests that even brief rater training sessions can result in ratings with reasonably good psychometric characteristics. For example, in research that employed 5-15 minutes of rater training, mean ratings have been between 5 and 6 on a 9-point scale, with standard deviations between 1.25 and 2.00. In addition, interpretable factor analyses have resulted, suggesting that halo was not overly severe, and interrater reliabilities have been in the .55 - .85 range (e.g., Borman, Rosse, Abrahams, & Toquam, 1979; Hough, 1984; Peterson & Houston, 1980).

Thus, as a starting point, the rater training program we have developed and revised over the past several years was adapted for use in this project. Components of the rater orientation and training program are described below.

1. Rater selection guidelines were prepared, as already described. Where feasible, two supervisors and four peers were identified as raters for each first-tour soldier ratee. To be eligible to rate a soldier, the supervisor or peer must be familiar with the ratee's performance and have supervised or worked with the ratee for at least one to two months.

2. A briefing was prepared to acquaint participant raters with the main objectives of Project A and to explain where the performance ratings fit into the project. As part of this briefing, the points about confidentiality of the ratings and the for-research-only nature of the ratings were emphasized. Also, a strong plea was made for raters to help us with this project and to do their best to provide accurate performance judgments.

3. An orientation to the behavior-based rating scales was developed. The idea here was to instruct raters on how to use the behavioral anchors systematically to make relatively objective performance ratings. In this part of the program, the principle of matching observed ratee performance with performance described in the scales' behavioral anchors was carefully explained and illustrated with several hypothetical examples. Emphasis was on getting raters to read each behavioral example on each scale and then to perform the matching step so that every rater was using the same rating standards (i.e., the behavioral anchors).

4. We prepared a short training program on three common rating errors -- halo, stereotyping, and too much attention paid to one or two events relevant to the ratee's performance (heretofore labeled as "one-incident-of-performance error"). Halo was explained and illustrated, and raters were asked to

BEST COPY AVAILABLE

acknowledge each ratee's strengths and weaknesses in their ratings. In the stereotyping discussion, the point was made that raters should consider only performance-related information directly relevant to the rating category and not information such as a ratee's education level, family background, or previous work experience. Finally, the trainer instructed raters to evaluate performance over time, not just a single outstanding or poor performance.

5. Peer raters were asked to make self-ratings using the Army-wide behavior-based rating scales, primarily for warm-up and practice, to ensure they became acquainted with the rating process before they began their evaluations. As mentioned, peer raters typically have no experience evaluating performance, and we felt that providing them with a self-rating warm-up using our scales would facilitate their rating task.

In sum, the main features of the orientation and training program were: First, the program was short (approximately 15 minutes) to increase the likelihood of holding participants' attention, while getting all of the important information points across; second, the points were made very simply during all parts of the program; third, the briefing covered all important points related to using the scales; fourth, confidentiality and the for-research-only purpose of the ratings were emphasized; and fifth, raters were urged to make accurate performance ratings.

The orientation and training program, as described here, was developed for the Batch A field tests. The idea was to start with this program, evaluate its effectiveness in the Batch A tests, revise for Batch B based on Batch A experience, continue the tryout in Batch B, and finally, revise as required for the large-scale Concurrent Validation effort. The program revisions are described in the following sections.

Lessons Learned During the Batch A Field Tests

The rater training and orientation program described above was used to train raters during the Batch A field tests; in addition, peers were given a self-rating task to familiarize them with the rating scales and provide practice in making ratings prior to evaluating their co-workers.

The Batch A rater training and orientation program seemed quite successful in that: (1) it appeared to flow well and be acceptable to both peer and supervisor raters; (2) the interrater reliabilities were very respectable, especially in light of the fact that the peer raters were inexperienced evaluators of performance; (3) the rating distributions were reasonable, with no extreme skew; (4) the effects of the training did not seem to be trainer-bound in that at least seven trainers administered the program at one time or another during the Batch A field tests; and (5) the relationships between the ratings

BEST COPY AVAILABLE

and other criterion variables showed some predictable patterns (e.g., correlations between MOS task scale ratings and hands-on test scores averaged about .25). All of the actual data analysis results relevant to the ratings are presented in Chapter 4 of this report.

Although the Batch A orientation and training program seemed effective, our experience during the field tests suggested various additions and modifications for further improving the program. First, the major substantive change involved inclusion of training for a fourth rating error. During the first field test at Ft. Polk, we observed that some supervisor and peer raters were evaluating all of their ratees at approximately the same level of effectiveness on many of the dimensions. Because it was important that raters make proper discriminations between ratees, we added a training component that would promote such discriminations where they were appropriate. Specifically, raters were encouraged not only to tell us about each individual's strengths and weaknesses (thereby avoiding halo error) but to also indicate differences between soldiers who perform well in a particular rating category and those who perform less well in the category. These instructions proved successful in that we observed fewer ratings at the same effectiveness level within categories subsequent to including this error-training component in the program.

Although we believe strongly that error reduction training is very important in yielding high quality evaluations, recent research (McIntyre, Smith, & Hassett, 1984; Pulakos, 1984) has suggested that error training alone may be insufficient for increasing rating accuracy, which is the crucial criterion for evaluating performance rating quality (Ilgen & Feldman, 1983; Landy & Farr, 1980). Therefore, subsequent to the Batch A field tests, we incorporated a more comprehensive accuracy training component into the program. That is, we stressed the notion that although we did not want raters to make rating errors, most important is that they rate each of their subordinates or co-workers accurately. Thus, if raters felt that their ratees actually performed at the same effectiveness level in a given performance category or that a particular soldier performed at approximately the same level across several categories, then they were encouraged to rate those individuals in this way. However, it was also emphasized that when real differences exist, the ratings should reflect these differences.

Finally, following the Batch A field tests a question was raised regarding the usefulness of self-ratings as an aid in familiarizing raters with the rating scales. This issue was especially relevant because less time was to be available for ratings during Concurrent Validation. It was thus important to consider which instruments and/or which aspects of the training program might be eliminated for Concurrent Validation.

Toward this end, we believed an empirical evaluation of the self-rating effects on the ratings was in order. If, for example, this portion of the program could be excluded, valuable testing time would be available for another purpose. It is also important to note that

BEST COPY AVAILABLE

no research to date has investigated the effects of self-ratings on subsequent ratings of others; thus, evaluating this aspect of the training had scientific as well as practical merit. Accordingly, an experiment was designed to investigate the self-rating effects. This experiment, which was conducted as part of the Batch B field tests, is described in the following chapter.

To summarize, two additions were made to the rater orientation and training program subsequent to the Batch A field tests. These were: (1) inclusion of training for the same-level-of-effectiveness error; and (2) an expanded discussion of rating accuracy concepts. New training scripts were written including these changes, and the revised program was implemented and pilot tested during the Batch B field tests. Our experiences with this program and what we learned during the field tests are described in the following section.

Lessons Learned During the Batch B Field Tests

The revised rater orientation and training program used during the Batch B field tests seemed quite successful. Both supervisor and peer raters were generally attentive and appeared to complete their ratings responsibly. The interrater reliabilities and rating distributions were very respectable (see Chapter 4 of this report). And again, the program did not seem to be trainer-bound in that both experienced administrators and less experienced staff members, who had been trained as rating program administrators, conducted the rating sessions.

To obtain the best possible program, we asked that each Batch B rating session administrator provide written feedback on his or her experiences with the training, outlining any suggestions for improving the program. Based upon this feedback, it was concluded that no major changes were required. However, several suggestions were made to facilitate administering the program for Concurrent Validation and to prevent errors in completing the rating forms.

First, one comment was that, rather than merely discussing the rating errors, it might be beneficial to actually show trainees what, for example, halo error and same-level-of-effectiveness error "look like." To this end, we enlarged one page of the Army-wide performance rating dimensions to poster size and then completed these scales to demonstrate the two errors (see Figures 2 and 3 for our depictions of halo and same-level-of-effectiveness errors, respectively). Unfortunately, stereotyping and one-incident-of-performance error could not be shown pictorially. In addition to enhancing understanding of two of the rating errors, these visual aids also proved useful for explaining the rating scale characteristics (e.g., the behavioral anchors, how to rate up to five soldiers using only one rating booklet) to trainees.

BEST COPY AVAILABLE

CATEGORY A: TECHNICAL KNOWLEDGE/SKILL

How effective is each soldier in displaying job and soldiering knowledge/skill?

		DOES NOT DISPLAY THE KNOWLEDGE/SKILL REQUIRED TO PERFORM MANY JOB ASSIGNMENTS AND TASKS.		DISPLAYS THE KNOWLEDGE/SKILL REQUIRED TO PERFORM MOST JOB ASSIGNMENTS AND TASKS PROPERLY, BUT MAY NEED HELP FOR HARDER TASKS.			DISPLAYS THE KNOWLEDGE/ SKILL TO PERFORM ALL JOB ASSIGNMENTS AND TASKS PROPERLY.	
Line up the names	1	①	②	③	④	●	⑥	⑦
of the soldiers	2	①	②	③	④	⑤	⑥	⑦
you are rating	3	①	②	③	④	⑤	⑥	⑦
with the rows	4	①	②	③	④	⑤	⑥	⑦
to the right.	5	①	②	③	④	⑤	⑥	⑦

CATEGORY B: EFFORT

How effective is each soldier in showing extra effort on the job/mission/assignment?

		DOES NOT PUT IN THE EFFORT TO MAKE SURE THE JOB GETS DONE; MAY GIVE UP EASILY WHEN FACED WITH DIFFICULT PROBLEMS OR SITUATIONS.		PUTS IN THE EXTRA EFFORT AND KEEPS TRYING WHEN IT'S VERY IMPORTANT TO COMPLETE ASSIGNMENTS; OVERCOMES OBSTACLES/ADVERSITIES ON THE JOB, IN GARRISON AND IN THE FIELD.			OFTEN VOLUNTEERS TO WORK EXTRA HOURS AND PUSHES ON TO OVERCOME ALL DIFFICULTIES AND ADVERSITIES UNTIL THE JOB IS DONE.	
Line up the names	1	①	②	③	④	●	⑥	⑦
of the soldiers	2	①	②	③	④	⑤	⑥	⑦
you are rating	3	①	②	③	④	⑤	⑥	⑦
with the rows	4	①	②	③	④	⑤	⑥	⑦
to the right.	5	①	②	③	④	⑤	⑥	⑦

CATEGORY C: FOLLOWING REGULATIONS AND ORDERS

How effective is each soldier in adhering to regulations, orders, and SOP and displaying respect for superiors?

		OFTEN FAILS TO FOLLOW ARMY/UNIT RULES, REGULATIONS, OR ORDERS; MAY SHOW DISRESPECT TOWARD SUPERIORS.		ALMOST ALWAYS FOLLOWS ARMY/UNIT RULES AND REGULATIONS; ALWAYS OBEYS ORDERS.			CAREFULLY FOLLOWS THE SPIRIT AND LETTER OF ARMY/UNIT RULES AND REGULATIONS; OBEYS ORDERS QUICKLY AND WITH ENTHUSIASM.	
Line up the names	1	①	②	③	④	●	⑥	⑦
of the soldiers	2	①	②	③	④	⑤	⑥	⑦
you are rating	3	①	②	③	④	⑤	⑥	⑦
with the rows	4	①	②	③	④	⑤	⑥	⑦
to the right.	5	①	②	③	④	⑤	⑥	⑦

CATEGORY D: INTEGRITY

How effective is each soldier in displaying honesty and integrity in job-related and personal matters?

		MAKES UP EXCUSES TO AVOID DUTY/ASSIGNMENTS, FAILS TO TAKE RESPONSIBILITY FOR ANY JOB-RELATED MISTAKES; MAY BE UNTRUTHFUL ABOUT JOB OR PERSONAL MATTERS.		ADMITS AND TAKES RESPONSIBILITY FOR MOST JOB-RELATED MISTAKES HE/SHE MAKES; IS TRUTHFUL WHEN QUESTIONED ABOUT JOB OR PERSONAL MATTERS.			TAKES EXTRA STEPS TO ENSURE THAT OTHERS ARE NOT BLAMED FOR HIS/HER MISTAKES; IS ALWAYS HONEST, EVEN WHEN IT MAY GO AGAINST PERSONAL INTERESTS.	
Line up the names	1	①	②	③	④	●	⑥	⑦
of the soldiers	2	①	②	③	④	⑤	⑥	⑦
you are rating	3	①	②	③	④	⑤	⑥	⑦
with the rows	4	①	②	③	④	⑤	⑥	⑦
to the right.	5	①	②	③	④	⑤	⑥	⑦

Figure 2. Pictorial Depiction of Halo Error

BEST COPY AVAILABLE

CATEGORY A: TECHNICAL KNOWLEDGE/SKILL

How effective is each soldier in displaying job and soldiering knowledge/skill?

		DOES NOT DISPLAY THE KNOWLEDGE/SKILL REQUIRED TO PERFORM MANY JOB ASSIGNMENTS AND TASKS.		DISPLAYS THE KNOWLEDGE/SKILL REQUIRED TO PERFORM MOST JOB ASSIGNMENTS AND TASKS PROPERLY, BUT MAY NEED HELP FOR HARDER TASKS.			DISPLAYS THE KNOWLEDGE/SKILL TO PERFORM ALL JOB ASSIGNMENTS AND TASKS PROPERLY.	
Line up the names	1	①	●	③	④	⑤	⑥	⑦
of the soldiers	2	①	●	③	④	⑤	⑥	⑦
you are rating	3	①	●	③	④	⑤	⑥	⑦
with the rows	4	①	●	③	④	⑤	⑥	⑦
to the right.	5	①	②	③	④	⑤	⑥	⑦

CATEGORY B: EFFORT

How effective is each soldier in showing extra effort on the job/mission/assignment?

		DOES NOT PUT IN THE EFFORT TO MAKE SURE THE JOB GETS DONE; MAY GIVE UP EASILY WHEN FACED WITH DIFFICULT PROBLEMS OR SITUATIONS.		PUTS IN THE EXTRA EFFORT AND KEEPS TRYING WHEN IT'S VERY IMPORTANT TO COMPLETE ASSIGNMENTS; OVERCOMES OBSTACLES/ADVERSITIES ON THE JOB, IN GARRISON AND IN THE FIELD.			OFTEN VOLUNTEERS TO WORK EXTRA HOURS AND PUSHES ON TO OVERCOME ALL DIFFICULTIES AND ADVERSITIES UNTIL THE JOB IS DONE.	
Line up the names	1	①	②	③	④	⑤	●	⑦
of the soldiers	2	①	②	③	④	⑤	●	⑦
you are rating	3	①	②	③	④	⑤	●	⑦
with the rows	4	①	②	③	④	⑤	●	⑦
to the right.	5	①	②	③	④	⑤	⑥	⑦

CATEGORY C: FOLLOWING REGULATIONS AND ORDERS

How effective is each soldier in adhering to regulations, orders, and SOP and displaying respect for superiors?

		OFTEN FAILS TO FOLLOW ARMY/UNIT RULES, REGULATIONS, OR ORDERS; MAY SHOW DISRESPECT TOWARD SUPERIORS.		ALMOST ALWAYS FOLLOWS ARMY/UNIT RULES AND REGULATIONS; ALWAYS OBEYS ORDERS.			CAREFULLY FOLLOWS THE SPIRIT AND LETTER OF ARMY/UNIT RULES AND REGULATIONS; OBEYS ORDERS QUICKLY AND WITH ENTHUSIASM.	
Line up the names	1	①	②	③	●	⑤	⑥	⑦
of the soldiers	2	①	②	③	●	⑤	⑥	⑦
you are rating	3	①	②	③	●	⑤	⑥	⑦
with the rows	4	①	②	③	●	⑤	⑥	⑦
to the right.	5	①	②	③	④	⑤	⑥	⑦

CATEGORY D: INTEGRITY

How effective is each soldier in displaying honesty and integrity in job-related and personal matters?

		MAKES UP EXCUSES TO AVOID DUTY/ASSIGNMENTS; FAILS TO TAKE RESPONSIBILITY FOR ANY JOB-RELATED MISTAKES. MAY BE UNTRUTHFUL ABOUT JOB OR PERSONAL MATTERS.		ADMITS AND TAKES RESPONSIBILITY FOR MOST JOB-RELATED MISTAKES HE/SHE MAKES. IS TRUTHFUL WHEN QUESTIONED ABOUT JOB OR PERSONAL MATTERS.			TAKES EXTRA STEPS TO ENSURE THAT OTHERS ARE NOT BLAMED FOR HIS/HER MISTAKES; IS ALWAYS HONEST, EVEN WHEN IT MAY GO AGAINST PERSONAL INTERESTS.	
Line up the names	1	①	②	●	④	⑤	⑥	⑦
of the soldiers	2	①	②	●	④	⑤	⑥	⑦
you are rating	3	①	②	●	④	⑤	⑥	⑦
with the rows	4	①	②	●	④	⑤	⑥	⑦
to the right.	5	①	②	③	④	⑤	⑥	⑦

Figure 3. Pictorial Depiction of Same-Level-of-Effectiveness Error

BEST COPY AVAILABLE

A second suggestion was to provide trainees with very explicit verbal instructions on how to complete machine-scannable instruments in general, and our rating forms in particular. For example, to ensure that the rating data could be matched with other measures for data analyses, it was critical that raters take great care to complete those sections of the forms that would be used for this matching process.

For example, since some data were to be matched using social security numbers (SSNs), it was essential that raters accurately complete the machine-scannable grids for their SSNs. A section emphasizing the importance of this step therefore was included in the orientation portion of the rater training program. Further, rather than asking raters to grid their SSNs on every form (the procedure during the Batch A and B field tests), it was decided that raters would do this on only one form, thereby reducing the chances of error, and all of a rater's completed instruments were then packaged together. Thus, data for an individual were matched with his or her other data solely on the basis of one accurately recorded SSN.

Raters also completed a Rating Assignment Form on which they were asked to record a three-digit code that had been assigned to each of their ratees. These codes would be used to match peer and supervisor ratings of each soldier to that soldier's predictor and other criterion data. To facilitate accuracy in recording ratee codes, step-by-step instructions for completing the Rating Assignment Forms were included in the orientation and training program.

Steps were also taken to prevent several other types of errors that we observed raters make in completing the forms during the field tests. In addition to providing verbal instructions for each instrument, rating session administrators were required to perform several data verification checks. This involved checking every completed instrument to ensure that it had been filled out completely and correctly. Administrators were alerted to potential problems and mistakes we had encountered in the past and told to check specifically for these. A list of the critical data verification checks that was distributed to administrators appears in Figure 4.

During the Batch B field tests, the administrator had provided a brief refresher of the error training points prior to completing each new instrument. We observed that refresher training prior to each new instrument seemed to be overly redundant, as this involved reviewing the points at least five times during the three-hour session. It was decided that refresher training probably would be more effective if it were provided only once during the program.

Finally, because many personnel would be administering ratings during Concurrent Validation, it seemed important to provide explicit instructions on how to run all aspects of the rating sessions. Therefore, a comprehensive rating session administration manual was written that included not only detailed rater orientation and training program scripts but also specific directions for making the peer

BEST COPY AVAILABLE

- Name and SSN must be written legibly on all forms and booklets.
- "Rating Group" must be completed on each booklet where this applies.
- SSN grids must be filled in correctly on the Background Information Form.
- MOS must be filled in correctly on the Background Information Form.
- Ratee code numbers must be filled in correctly on the Ratee Assignment Form (see page 1 of Form 6A). Also, for each ratee, the box number on the form must correspond to the line number on which his or her name appears on the Ratee Name Tab.
- The two questions that appear under the Ratee Assignment Form (Form 6A) must be answered for each ratee.
- All grids containing responses must be filled in completely with dark marks.
- The lines used under each rating scale must correspond to the lines that contain names on the Ratee Name Tabs. If a rater is evaluating three people, only the first three ratee lines should be used for each item.
- All items on each form should be completed. The only exception is if a rater indicates that he or she cannot rate a soldier in some area(s) and the scale does <u>not</u> contain a "Not Observed" option. Always encourage the rater to provide a rating if she or he possibly can. If, however, the individual insists that he or she has no idea how the ratee performs in the area(s), instruct the rater to leave the item(s) blank.
- Check for halo error and same-level-of-effectiveness error. Encourage raters to make distinctions where they can.
- Completed <u>machine-scannable</u> forms are to be placed in each rater's confidential envelope. <u>One</u> envelope should be used to collect all of the forms for a given rater.

Figure 4. Critical Data Verification Checks: Rating Sessions

BEST COPY AVAILABLE

assignments, obtaining the correct supervisor raters, responding to common questions that arise regarding the ratings, and dealing with potential problems. This manual appears in Appendix I.

In addition, plans were made to have all Concurrent Validation rating session administrators participate in a two and one-half day training workshop that would include instruction, practice, and feedback on running the rating sessions.

In summary, then, no major changes were made to the training program as a result of our Batch B field experience. However, the program was expanded considerably to include explicit administrative instructions for Concurrent Validation, related both to running standardized, effective rating sessions and to ensuring that high quality data were collected.

BEST COPY AVAILABLE

CHAPTER 6: BATCH B RATER TRAINING EXPERIMENT: THE EFFECTS OF PRACTICE ON MAKING RATINGS

Elaine D. Pulakos

Purpose of the Experiment

As discussed in Chapter 5, the purpose of the training experiment conducted during the Batch B field tests was to determine whether providing peer raters with practice significantly improved performance rating quality beyond what was obtained by the rater orientation and training program alone. Two training treatments were evaluated for the peers: (1) rater orientation and error reduction training, including a brief refresher of the error training points prior to administering each new scale, and (2) this same program plus a self-rating warm-up for each scale.

Parallel training treatments were also developed and evaluated for the supervisors. However, because the rating scales had been specifically developed to evaluate first-term soldier performance, having the supervisors perform a self-rating task using these scales would have been inappropriate. Instead, practice for the supervisors entailed rating a description of one hypothetical soldier prior to evaluating their subordinates. The two supervisor training treatments were: (1) rater orientation and error reduction training, including brief refresher training before each new instrument, and (2) this same program plus a practice rating of one hypothetical soldier on six Army-wide behavioral dimensions.

The training treatments for each peer and supervisor rater group were evaluated in terms of their effects on rating accuracy and three rating errors (halo, leniency/severity, and restriction-of-range).

Method

Subjects

A total of 817 peer raters and 660 supervisor raters participated in the Batch B field tests. Each soldier represented one of the following five MOS: 11B (Infantryman), 19E (Armor Crewman), 31C (Single Channel Radio Operator), 63B (Light Wheeled Vehicle Mechanic), and 91A (Medical Specialist). Data were collected from four CONUS locations (Fort Stewart, Fort Lewis, Fort Riley, and Fort Polk) and USAREUR. Table 11 shows the breakdown of peer and supervisor raters from each location by MOS.

BEST COPY AVAILABLE

Table 11

Breakdown of the Rater Training Sample by Location and MOS

Rater Group	USAREUR	Fort Stewart	Fort Lewis	Fort Riley	Fort Polk	Totals
11B Incumbents	58	31	19	30	30	168
Supervisors	54	26	18	20	18	136
19E Incumbents	57	30	30	24	31	172
Supervisors	59	24	21	16	28	148
31C Incumbents	58	23	16	26	26	149
Supervisors	55	15	16	26	13	125
63B Incumbents	61	27	13	29	26	156
Supervisors	75	19	18	33	20	165
91A Incumbents	64	21	23	34	30	172
Supervisors	54	11	11	18	12	86
TOTALS	298	132	101	143	143	817
	277	95	84	113	91	660

BEST COPY AVAILABLE

Procedure and Design

First-term soldiers reported to their rating sessions in groups of approximately 15. At each location, only one supervisor rating session was conducted for each MOS. Thus, it was necessary to assign raters within an MOS at a particular post to one of the two training treatments and then counterbalance the treatment for each MOS across the posts. So, for example, at Fort Stewart, 19E and 91A peers and supervisors received error training only, while 11B, 31C, and 63B peers and supervisors received error training plus practice. Conversely, at Fort Lewis, 11Bs, 31Cs, and 63Bs received error training only, while 19Es and 91As received error training plus practice. A similar counterbalancing scheme was used for the remaining three locations.

It is important to note, however, that data were collected from twice as many soldiers in USAREUR as were collected at any single CONUS location. Thus, the assignment process described above resulted in approximately equal numbers of soldiers from each MOS receiving each type of training across the five data collection sites.

Rating Instruments

Four of the rating instruments used during the Batch B field tests were relevant for the present study. Specific details regarding the development of each scale type appear elsewhere (see Chapter 2 of this report for the Army-wide scales and Toquam et al., 1985, for the MOS-specific scales). The instruments were:

1. Army-wide behavioral rating scales. These scales consisted of 11 Army-wide behavior-based rating dimensions (e.g., Effort; Following Regulations and Orders), representing aspects of overall soldier effectiveness that are relevant to all MOS.

2. Army-wide common task scales. These scales consisted of 12 Army-wide common task dimensions (e.g., Identify Threat; Use a Map), employing no behavioral anchors and a "Not Observed" option. The task areas included in these scales were derived from the Skill Level 1 Common Task Soldiers' Manual.

3. MOS-specific behavioral rating scales. These scales consisted of 6-12 behavior-based rating dimensions (e.g., Vehicle and Equipment Operation - 63B; Avoiding Enemy Detection - 11B) relevant to job performance in a given MOS. Hence, there were five versions of the instrument, one for each of the Batch B MOS.

BEST COPY AVAILABLE

4. MOS-specific task scales. These scales consisted of 15 MOS-specific task dimensions (e.g., Establish, Enter, or Leave a Radio Net - 31C; Initiate an Intraveneous Infusion - 91A), employing no behavioral anchors and a "Not Observed" option. There were also five different versions of this instrument, one for each target MOS. Each set of scales was matched directly to the 15 tasks tested hands-on for the MOS.

Training Programs

Rater Orientation and Error Training Only

Peer and supervisor raters assigned to this experimental condition (EO) received training that can be characterized as a combination psychometric error and frame-of-reference program (Bernardin & Pence, 1981; Pulakos, 1984). The philosophy behind this training is outlined in Chapter 5 of this report.

Briefly, one component of training involved carefully explaining the logic of the behavior-based and task rating scales as well as urging raters to study and properly use the instruments to arrive at their evaluations. The second major component of training involved description of halo, stereotyping, one-incident-of-performance, and same-level-of-effectiveness errors in lay terms and provided guidance on how to avoid these errors.

Rater Orientation and Error Training Plus Practice: Peer Raters

This experimental condition (E+P) consisted of the same training outlined above plus practice using the rating scales in the form of self-appraisals. Specifically, prior to rating their co-workers on each of the four sets of scales (the Army-wide behavioral rating scales, the Army-wide common task scales, the MOS-specific behavioral rating scales, and the MOS-specific task scales), peer raters were asked to evaluate themselves using these instruments.

Rater Orientation and Error Training Plus Practice: Supervisor Raters

Supervisors assigned to this training condition also received the rater orientation and error reduction training discussed above. However, practice for the supervisors entailed evaluating one hypothetical ratee on the Army-wide behavioral performance dimensions. A vignette (see Figure 5) describing performance of a first-term soldier was developed for this purpose. The behavioral examples used in the vignette were obtained from the pool of critical incidents retranslated during Army-wide behavior scale development.

BEST COPY AVAILABLE

Soldier Description

Mike Bennett has been in the Army for nine months. When he first enlisted, he was not planning to make a career there. However, good things have happened to Mike since he entered, and now he thinks he may stick around awhile. Mike has really enjoyed learning about various "field techniques," and he spends his own time practicing his individual soldier skills. His interest in these areas has paid off. For example, when he was recently instructed to negotiate a land navigation course, he ran the entire course, hitting every check point, in record time. On another occasion, he thoroughly instructed his men on the escape and evasion exercise. Consequently, during the course of the actual exercise, all members of the squad made it through and set a time record for running the escape and evasion course.

Mike is very knowledgeable about various pieces of equipment and also takes pride in the equipment he is responsible for. One time, for instance, when Mike was tasked with repairing a tank, he stayed up all night working on it until the tank was fully operational. He has consistently maintained his equipment and, over a period of seven months, it was not deadlined once. In fact, Mike spends so much time working on equipment, that he maintains two sets of uniforms: one inspection ready and one for work. He changes from a dirty uniform into a good one everytime he leaves AO. However, sometimes Mike gets so involved fixing equipment that he is late for other activities. In the past six months, for example, Mike has reported late for duty three times, although he was never more than about five minutes late and his tardiness never occurred under circumstances that would adversely affect the unit.

One thing Mike is not really wild about is "clean-up" work. One day, especially, Mike just was not at all motivated to do his cleaning tasks. When asked, he told his commander that he had completed his tasks for inspection, although he had really not done anything. However, Mike made up for this at the next inspection by working on his yard and sidewalk even after everyone else had given up. As a result, he won the yard of the month award and the Best Looking Quarters.

Mike has never been particularly athletic. When he first enlisted, he was unsure of whether or not he would be able to deal with all the PT. Once he realized he might want to reenlist, though, he decided he had better get in shape. So, after failing his first PT test, he began working out more and was able to pass his second PT test.

Mike has also enjoyed the people he has met in the Army and spends considerable time hanging around with his buddies. Typically, they will go to the bar, have a few beers, and generally have a great time. One night, however, Mike had too much to drink. He got involved in a large fist fight at a local club, and caused damage to the premises. Luckily, his friends helped to break up the fight before anyone was seriously injured.

Figure 5. Supervisor Practice Vignette

BEST COPY AVAILABLE

Development of Vignettes to Assess Accuracy

It has been argued that accuracy is the crucial criterion for assessing the quality of performance ratings (e.g., Borman, 1979). Evaluating psychometric indices such as leniency or halo is useful for identifying errors in ratings, but such indices must be considered as indirect measures of accuracy. Of course, assessing the accuracy of ratings requires that the actual performance levels of ratees be known, and this requirement is almost always impossible to achieve.

In the present research we were able to create "ratees" with known performance scores by developing vignettes about first-term soldiers performing their jobs, and using in the vignettes previously scaled behavioral examples (just as we did for the supervisor practice rating condition). The true or target performance level for a dimension was simply the mean retranslation effectiveness level for the example included in the vignette for that dimension.

A total of four vignettes were written describing performance in the following six Army-wide areas: (1) Effort; (2) Maintaining Assigned Equipment; (3) Maintaining Living and Work Areas; (4) Physical Fitness; (5) Self-Development; and (6) Self-Control. The specific procedures used to develop the vignettes are described below.

By using expert judges' estimates of the true intercorrelations between the six dimensions along with dimension means of 4.0 and standard deviations of 1.5, a true score matrix (see Table 12) containing scores for hypothetical ratees on each dimension was generated. This matrix possessed the "correct" covariance structure.

Using behavioral examples obtained from the retranslation phase of the Army-wide behavior scaling process, vignettes were then written describing four ratees performing at the effectiveness levels shown in Table 12. It is important to note that in the retranslation process each incident had been allocated reliably into a single dimension and assigned a narrow range of effectiveness levels.

After evaluating their co-workers or subordinates on the Army-wide behavioral rating scales, both peers and supervisors read and then rated the soldiers described in each of the four vignettes. The materials used to collect these data, including instructions, the actual vignettes, and special rating scales containing only the six dimensions relevant to the vignettes, are presented in Appendix J.

Dependent Variables

Using the peer and supervisor ratings of the soldiers evaluated (not the vignettes), the following four rating indices were computed: interrater agreement, halo, leniency/severity, and restriction-of-range. The vignette ratings were used to assess training effects on accuracy. Each dependent measure (described in detail below) was examined separately for peer and supervisor raters.

BEST COPY AVAILABLE

Table 12

True Score Matrix for Vignette Ratees on Six Army-Wide Dimensions

Dimensions	Ratees			
	1	2	3	4
1. Effort	5.0	2.0	6.0	4.0
2. Maint Assign Equip	5.0	3.0	5.0	2.0
3. Maint Living & Work Areas	3.0	1.0	5.0	3.0
4. Physical Fitness	4.0	3.0	5.0	6.0
5. Self-Development	7.0	2.0	6.0	4.0
6. Self-Control	6.0	1.0	2.0	5.0

Note. Because the rating task required evaluators to select a whole number from 1 to 7 describing each soldier's effectiveness on a dimension, the generated true scores were rounded to the nearest whole number.

BEST COPY AVAILABLE

Interrater Agreement

Within each training treatment, an intraclass correlation coefficient was computed for each behavioral or task dimension of the four rating instruments on which peer and supervisor ratings of soldier performance were obtained. Within each instrument, these correlations were then averaged across the dimensions, resulting in four indices of rater agreement for each training condition.

Rating Errors

Using the behaviorally based ratings of Army-wide and MOS-specific performance, three rating errors were examined. It is important to note, however, that each of the error measures was calculated at the treatment group level of analysis (i.e., across raters and ratees) rather than calculating an error measure for each individual rater. The latter was not possible because many raters, especially the supervisors, evaluated only one or two ratees. In order to obtain relatively stable estimates of individual rater errors, each rater would have had to evaluate a minimum of four ratees. Because this was not the case, computation of the errors proceeded as follows. First, for each dimension, all raters' ratings of a given ratee were averaged. Thus, the n size for all analyses was the number of ratees. Three error indices were then computed for each the Army-wide behavioral dimensions and the MOS-specific behavioral dimensions as follows.

Halo. Separately for the Army-wide and the MOS-specific scales, an average interdimension correlation was computed across ratees within each of the two training conditions.

Leniency/Severity. Dimension means were computed across ratees within each training treatment. These means were then averaged across the dimensions within each rating instrument and training treatment.

Restriction-of-Range. Dimension variances were computed across ratees within each training treatment. These variances were averaged across the dimensions to provide two final measures of the error (one based on the Army-wide ratings, the other on the MOS-specific ratings) within each training condition.

Accuracy

The effects of training on rating accuracy were assessed using only the vignette rating data. However, because each rater evaluated four vignette ratees, it was possible to compute accuracy for each individual rater. Accuracy was operationalized as the average squared difference between each rater's ratings of the vignette ratees and the vignette true scores, with lower values indicating higher accuracy.

BEST COPY AVAILABLE

Results

Training Effects on Interrater Agreement

For the peer and supervisor rater groups, Table 13 contains the average intraclass correlations for each of the four rating instruments within training condition. To determine whether or not there were significant differences between the treatments, χ^2 tests were performed. Specifically, for each rating instrument and separately for the peers and supervisors, a χ^2 test was performed to assess whether or not there was a significant difference in interrater agreement difference between the group of raters who received error training only and the group of raters who received error training plus practice. The χ^2 tests performed here were, however, slightly different than typically encountered significance tests. Thus, prior to presenting the results of these tests, a brief discussion of the present procedures will be undertaken.

Conducting a significance test to determine whether or not there is a difference between two correlations requires knowing the standard error (hence, the variance) of the correlations being compared. Recall that the average intraclass correlations shown in Table 13 were obtained by averaging the intraclass correlations across dimensions within each training treatment. The variance of an average is a function of the variance of each component going into the average as well as the covariances between these componenets. Because the ICCs going into each average were computed using data from the same raters, they were not independent and, consequently, the covariances among them were greater than zero. Because a different number of raters were involved in the calculation of each ICC, however, a precise calculation of the value for these covariances using normal theory results (Elston, 1975) was not possible. Thus, we decided to bracket the variance of the average ICC between its minimum and maximum value.

For a Fisher z transformed intraclass correlation, the variance is given by 1/n-1.5, where n is the number of observations. The maximum variance of the average is obtained by assuming perfect dependence among the components, giving the variance of the average as the variance of any single component or:

$$\text{Maximum variance of the average correlation} = \frac{1}{n-1.5}$$

The minimum variance is obtained by assuming independence of the components divided by the number of components entering the average. Assuming equal n and an average based upon k correlations,

$$\text{Minimum variance of the average correlation} = \frac{1}{k(n-1.5)}$$

BEST COPY AVAILABLE

Table 13

Interrater Reliabilites by Training Condition Across All MOS

	A-W Scales		A-W Tasks		MOS Scales		MOS Tasks	
Rater Group	EO	E+P	EO	E+P	EO	E+P	EO	E+P
Peers	.26	.31	.18	.17	.16	.18	.11	.11
Supervisors	.32	.37	.21	.26	.30	.38[a]	.21	.34[a]

Note. EO = error training only; E+P = error training plus practice; these are one rater reliabilities calculated on the unadjusted ratings.

[a] Minimum χ^2 was nonsignificant, but maximum χ^2 significant.

BEST COPY AVAILABLE

When testing for a difference between the error training only and error training plus practice average ICCs, the maximum variance can be used to compute the minimum χ^2 test statistic, while the minimum variance can be used to compute the maximum test statistic. The actual χ^2 statistic (i.e., the test statistic that correctly accounts for the covariation among the correlations) lies somewhere between the minimum and maximum values. The formula for the χ^2 test used was:

$$\chi^2 = \frac{(z_1 - z_2)^2}{v_1 + v_2}$$

where: z_1, and z_2 are the Fisher z transformed average correlations, or

$$z_i = -1/2 \ln \frac{(1 + \bar{r}_i)}{(1 - \bar{r}_i)}$$

and: v_1 and v_2 are either the minimums or the maximums for the variances of z_1 and z_2 respectively (see above).

The minimum and maximum χ^2 statistics were referred to a theoretical χ^2 distribution on one degree of freedom. Significance resulted if the observed test statistic exceeded 3.84, the .05 critical value. Interpretation of the minimum and maximum χ^2s for any given comparison proceeded as follows:

- A significant minimum χ^2 indicated a definite difference between the error training only and the error training plus practice ICCs.
- A nonsignificant maximum χ^2 indicated no difference between the two ICCs.
- A nonsignificant minimum χ^2 but a significant maximum χ^2 indicated a possible difference between the two ICCs, but no definitive conclusions could be drawn.

As shown in Table 13, there were no differences between the two training treatments for the peers on any of the rating scale types. For the supervisors, interrater agreement was consistent across the training treatments for the Army-wide scales. However, it appears that practice may have increased rater agreement on the MOS-specific scales.

Given that the supervisors' practice was restricted to only the Army-wide rating dimensions, the finding that practice seemed to facilitate agreement on the MOS-specific scales but <u>not</u> the Army-wide scales seemed counterintuitive. Hence, the data were inspected further to evaluate the consistency of this effect across MOS. These analyses

BEST COPY AVAILABLE

revealed a significant difference between the two training treatments on the MOS scales for only the 91As; there were no differences in interrater agreement as a result of training for any of the other MOS.

Training Effects on the Rating Errors

Tables 14 and 15 contain results for the rating error dependent variables by experimental condition for the Army-wide behavioral rating scales and the MOS-specific behavioral rating scales, respectively. Separately for the peers and supervisors, minimum and maximum χ^2s (like those described above) were computed to test for possible halo differences between the error training only and error training plus practice treatment conditions. F-tests were used to make similar comparisons for the leniency/severity and restriction-of-range errors. For both the peer and supervisor raters, there were no differences between the two training treatments in terms of halo, leniency/severity, or restriction-of-range on either set of rating scales.

Training Effects on Accuracy

A 2x2 (training x rater group) fixed-factor analysis of variance (ANOVA) was conducted to evaluate training effects on accuracy. Recall from the method section that accuracy was operationalized as the average squared difference between the vignette true scores and each rater's observed ratings of the four vignette ratees, with lower values indicating greater accuracy. The results of the ANOVA revealed no significant main effects for training, $F(1,1316) = .24$, ns., or rater group, $F(1,1316) = .03$, ns. The training x rating group interaction was also nonsignificant, $F(1,1316) = 2.35$, ns. The means for each condition appear in Table 16.

Summary and Conclusions

The purpose of this experiment was to assess whether a practice component of training improved performance rating quality beyond what was obtained by error training alone. Practice was provided to the peer raters by administering a self-rating task for all of the instruments prior to the peers' evaluating their co-workers. Supervisor raters received practice by evaluating one hypothetical ratee on six Army-wide behavioral dimensions prior to rating their subordinates.

Results of the study were identical for the peer and supervisor raters. The practice component of training yielded no significant improvement in ratings in terms of interrater agreement or any of the rating errors assessed here (i.e., halo, leniency/severity, and restriction-of-range). Further, practice did not facilitate accuracy on a vignette rating task. It was therefore concluded that the peer self-ratings and supervisor practice ratings should be eliminated from the rater orientation and training program for Concurrent Validation.

BEST COPY AVAILABLE

Table 14

Rating Error Dependent Variables by Training Condition and Rater Group: Army-Wide Behavioral Rating Scales

	Halo		Leniency/ Severity		Restriction-of-Range	
Rater Group	EO	E+P	EO	E+P	EO	E+P
Peers	.44	.36	4.42	4.60	1.29	1.10
Supervisors	.47	.45	4.55	4.64	1.74	1.68

Note. EO = error training only; E+P = error training plus practice; halo: average interdimension correlation; leniency/severity: average dimension mean; restriction-of-range: average dimension variance.

BEST COPY AVAILABLE

Table 15

Rating Error Dependent Variables by Training Condition and Rater Group: MOS-Specific Behavioral Rating Scales

	Halo		Leniency/ Severity		Restriction-of-Range	
Rater Group	EO	E+P	EO	E+P	EO	E+P
Peers	.46	.50	4.60	4.79	1.00	.85
Supervisors	.45	.55	4.78	4.73	1.19	1.35

Note. EO = error training only; E+P = error training plus practice; halo: average interdimension correlation; leniency/severity: average dimension mean; restriction-of-range: average dimension variance.

BEST COPY AVAILABLE

Table 16

Accuracy Means by Training Condition and Rater Group

Rater Group	Error Training	Error Training Plus Practice
Peers	1.60	1.75
	(.68)	(.89)
Supervisors	1.69	1.59
	(.92)	(.54)

Note. SDs appear in parentheses.

BEST COPY AVAILABLE

GENERAL SUMMARY AND CONCLUSIONS

The main objective of the Army-wide rating scale development effort was to create dimensions of soldier performance that would be relevant to first-term soldiers in any MOS. This is especially important for Project A, because relatively expensive MOS-specific performance measures could be developed for only nine of the 19 jobs included in the research. The Army-wide rating scales were thus intended to provide criterion measures for the remaining 10 MOS and any other MOS that might be involved in future personnel research projects with first-tour soldiers.

An initial conceptual model of soldier effectiveness helped guide empirical scale development efforts. The behavior analysis method was employed to identify soldier effectiveness dimensions and to develop behavioral definitions of performance in each dimension. The resulting Army-wide performance rating scales were then readied for field testing with supervisor and peer raters in nine different jobs.

In addition, the Skill Level I Common Task Soldier's Manual guided development of rating scales intended to measure performance in several task areas for which all first-tour soldiers are responsible. These scales were also readied for field testing. Finally, a rater orientation and training program was prepared to help supervisor and peer raters make their performance ratings as accurate as possible. This program was also tried out during field testing.

The research staff conducted two separate field tests of all performance measures, including hands-on and job knowledge tests, MOS-specific rating scales, and the Army-wide rating measures discussed in the present report. The first (Batch A) field test focused on four MOS, and the second (Batch B) field test focused on five other MOS. A total of 904 supervisor and 1,206 peer raters participated in the field tests, evaluating 1,369 first-term soldiers in all.

Results of the field tests were very encouraging. In particular: (1) rater participants seemed reasonably accepting of the rating program and appeared able to understand and comply with the instructions; (2) rating distributions were acceptable, with means a little above the scale mid-points and standard deviations comparable to those found in other research; and (3) interrater reliabilities were acceptably high, for both supervisor and peer raters.

Although results from both the Batch A and B field tests were on the whole positive, valuable information was gleaned from these trial administrations of the Army-wide scales. This experience guided our revisions of both the rating scales and the rater orientation and training program for Concurrent Validation.

COPY AVAILABLE

REFERENCES

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 60, 556-560.

Borman, W. C. (1979). Format and training effects on rating accuracy and rating errors. Journal of Applied Psychology, 64, 412-421.

Borman, W. C., Motowidlo, S. J., & Hanser, L. M. (1983, August). A model of individual performance effectiveness: Thoughts about expanding the criterion space. Paper in symposium, Integrated Criterion Measurement for Large Scale Computerized Selection and Classification, 91st Annual American Psychological Association Convention, Anaheim, CA.

Borman, W. C., Motowidlo, S. J., Rose, S. R., Hanser, L. M. (1985). Development of a model of soldier effectiveness. In ARI Technical Report 660, Improving the selection, classification, of Army enlisted personnel: Annual report, 1984 fiscal year. (AD A178 944). Alexandria, VA: Army Research Institute.

Borman, W. C., Rosse, R. L., Abrahams, N. M., & Toquam, J. L. (1979). Investigating personality and vocational interest constructs and their relationships with Navy recruiter performance. Minneapolis: Personnel Decisions Research Institute.

Borman, W. C., White, L. A., Gast, I. F. (1985, August). Factors relating to peer and supervisor ratings of job performance. Paper presented to the 92nd Annual American Psychological Association Convention, Toronto, Canada. In ARI Technical Report 660, Improving the selection, classification, of Army enlisted personnel: Annual report, 1984 fiscal year. (AD A178 944). Alexandria, VA: Army Research Institute.

Brown, E. M. (1968). Influence of training, method, and relationship on the halo effect. Journal of Applied Psychology, 52, 195-199.

Campbell, C. H., Campbell, R. C., Rumsey, M. G., & Edwards, D. C. (1986). Development and field test of task-based MOS-specific criterion measures. ARI Techical Report 717 (in press).

Campbell, J. P., Dunnette, M. D., Arvey, R., & Hellervik, L. (1973). The development and evaluation of behaviorally based rating scales. Journal of Applied Psychology, 57, 15-22.

BEST COPY AVAILABLE

Davis, R. H., Davis, G., & Joyner, J. (1985). Development and field testing of job relevant knowledge tests for selected MOS. ARI Technical Report (in preparation).

Elston, R. C. (1975). On the correlation between correlations. Biometrika, 62, 133-140.

Eaton, N. K., & Goer, M. H. (Eds.). (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Technical appendix to the annual report (ARI Research Note 83-37. (AD A137 117)

Eaton, N. K., Goer, M. H., Harris, J. H., & Zook, L. M. (Eds.). (1985). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1984 fiscal year (ARI Technical Report 660). (AD A178 944)

Hough, L. M. (1984). Development and evaluation of the "Accomplishment Record" method of selecting and promoting professionals. Journal of Applied Psychology, 69, 135-146.

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report (ARI Research Report 1347). (AD A141 807)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Project A - Research plan (ARI Research Report 1332). (AD A129 728)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute (1985). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report synopsis, 1984 fiscal year (ARI Research Report 1393). (AD A173 824)

Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, & Army Research Institute (1983). Improving the selection, classification, and utilization of Army enlisted personnel: Appendices to annual report, 1984 fiscal year (ARI Research Note 85-14).

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process approach. In B. M. Staw & L. L. Cummings (Eds.), Research in organizational behavior, Vol. 5, pp. 141-197.

Landy, F. J., & Farr, J. (1980). Performance rating. Psychological Bulletin, 87, 72-107.

BEST COPY AVAILABLE

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 60, 550-555.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purposes of rating. Journal of Applied Psychology, 69, 147-156.

Peterson, N. G. (Ed.). (1985). Development and field test of the trial battery for Project A. ARI Technical Report (in preparation).

Peterson, N. G., & Houston, J. S. (1980). The prediction of correctional officer job performance: Construct validation in an employment setting. Minneapolis, MN: Personnel Decisions Research Institute.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating formats. Organizational Behavior and Human Decisions Processes (in press).

Riegelhaupt, B. J., Harris, C. D., & Sadacca, R. (1985). The development of administrative measures as indicators of soldier effectiveness (draft).

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Toquam, J. L., McHenry, J. J., Corpe, V. A., Rose, S. R., Lammlein, S. E., Kemery, E., Borman, W. C., Mendel, R., & Bosshardt, M. (1985). Development and field test of behaviorally anchored rating scales for nine MOS. ARI Technical Report (in preparation).

Zedeck, S., & Cascio, W. (1982). Performance appraisal decisions as a function of rater training and purpose of the appraisal. Journal of Applied Psychology, 67, 752-758.

870615

BEST COPY AVAILABLE